On the limiting law of the length of the longest common and increasing subsequences in random words

Jean-Christophe Breton, Christian Houdré

On the limiting law of the length of the longest common and increasing subsequences in random words

Jean-Christophe Breton* Christian Houdré†

À la Mémoire de Marc Yor

Abstract

Let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two sequences of independent and identically distributed (iid) random variables taking their values, uniformly, in a common totally ordered finite alphabet. Let LCI_n be the length of the longest common and (weakly) increasing subsequence of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. As n grows without bound, and when properly centered and normalized, LCI_n is shown to converge, in distribution, towards a Brownian functional that we identify.

1 Introduction

We analyze below the asymptotic behavior of the length of the longest common subsequence in random words with an additional (weakly) increasing requirement. Although it has been studied from an algorithmic point of view in computer science, bio-informatics, or statistical physics, to name but a few fields, mathematical results for this hybrid problem are very sparse. To present our framework, let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two infinite sequences whose coordinates take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_m\}$, a finite totally ordered alphabet of cardinality m. Next, LCI_n, the length of the longest common and (weakly) increasing subsequences of the words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ is the maximal integer $k \in \{1, \ldots, n\}$, such that there exist $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$, satisfying the following two conditions:

*IRMAR, UMR 6625, Université de Rennes 1, 263 Avenue du Général Leclerc CS 74205, 35042, Rennes, France, jean-christophe.breton@univ-rennes1.fr. Many thanks to the School of Mathematics of the Georgia Institute of Technology for several visits during which part of this work was done.

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA, houdre@math.gatech.edu. Research supported in part by the grant #246283 from the Simons Foundation and by a Simons Foundation Fellowship grant #267336. Many thanks to the Centre Henri Lebesgue of the Université de Rennes 1, to the Département MAS of École Centrale Paris, to the LPMA of the Université Pierre et Marie Curie and to CIMAT, Gto, Mexico for their hospitality, where part of this work was done.

Keywords: Longest Common Subsequence, Longest Increasing Subsequence, Random Words, Random Matrices, Central Limit Theorem, Optimal Alignment, Last Passage Percolation.

MSC 2010: 05A05, 60C05, 60F05.

(i) $X_{i_s} = Y_{j_s}$, for all $s = 1, 2, \ldots, k$,

(ii) $X_{i_1} \le X_{i_2} \le \cdots \le X_{i_k}$ and $Y_{j_1} \le Y_{j_2} \le \cdots \le Y_{j_k}$.

LCI_n is a measure of the similarity/dissimilarity of the words often used in pattern matching, and its asymptotic behavior is the purpose of our study. (Asymptotically, the strictly increasing case is of little interest, having m as a pointwise limiting behavior.) For LCI_n, here is our result:

Theorem 1.1 *Let $X = (X_i)_{i \ge 1}$ and $Y = (Y_i)_{i \ge 1}$ be two sequences of iid random variables uniformly distributed on $\mathcal{A}_m = \{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_m\}$, a totally ordered finite alphabet of cardinality m. Let LCI_n be the length of the longest common and increasing subsequence of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. Then,*

$$\frac{LCI_n - n/m}{\sqrt{n/m}} \Longrightarrow \max_{0=t_0 \le t_1 \le \cdots \le t_{m-1} \le t_m = 1} \min\Bigg(-\frac{1}{m}\sum_{i=1}^m B_1^{(i)}(1) + \sum_{i=1}^m \Big(B_1^{(i)}(t_i) - B_1^{(i)}(t_{i-1})\Big), \\ -\frac{1}{m}\sum_{i=1}^m B_2^{(i)}(1) + \sum_{i=1}^m \Big(B_2^{(i)}(t_i) - B_2^{(i)}(t_{i-1})\Big)\Bigg), \tag{1.1}$$

where B_1 and B_2 are two m-dimensional standard Brownian motions on $[0, 1]$.

A first motivation for our work has its origins in the identification, first obtained by Kerov [Ker], of the limiting length (properly centered and scaled) of the longest increasing subsequence of a random word, as the maximal eigenvalue of a certain Gaussian random matrix. When combined with results of Baryshnikov [Bar] or Gravner, Tracy and Widom [GTW] (see also [BGH]), this limiting law has a representation as a Brownian functional. Moreover, [Ker, Chap. 3, Sec. 3.4, Theorem 2] showed that the whole normalized limiting shape of the RSK Young diagrams associated with the random word, is the spectrum of the traceless Gaussian Unitary Ensemble (GUE). Since the length of the top row of the diagrams is the length of the longest increasing subsequence of the random word, the maximal eigenvalue result is recovered. (The asymptotic length result was rediscovered by Tracy and Widom [TW] and the asymptotic shape one by Johansson [Joh]. Extensions to non-uniform letters were also obtained by Its, Tracy and Widom [ITW1, ITW2].) A second motivation for our work is the binary LCI_n result of [HLM], that we will revisit and extend, as well as the single-word results of [HL]. The dependence (or independence) structure between the two sequences X and Y is carried over into a similar structure between the two standard Brownian motions B_1 and B_2. Hence, when $X = Y$, our results recover, with the help of [BGH], the weak limits obtained in [Ker], [Joh], [TW], [ITW1], [ITW2], [HL], and [HX], while if X and Y are independent so are B_1 and B_2.

As for the content of the paper, the next section (Section 2) establishes a pathwise representation for the length of the longest common and increasing subsequence of the two words as a max/min functional. In Section 3, the probabilistic framework is initiated, the representation becomes the maximum over a random set of the minimum of random sums of randomly stopped random variables. The various random variables involved are studied and their (conditional) laws found. In Section 4, the limiting law is obtained. This is done in part by a derandomization procedure (of the random sums and of the random constraints) leading to the Brownian functional (1.1) of Theorem 1.1. In the last section (Section 5), various extensions and generalizations are discussed as well as some open questions related to this problem. Finally, Appendix A gives missing steps in the proof of the main theorem in [HLM] as well as corrections to arguments presented there; providing, in the much simpler binary case, a rather self-contained proof.

2 Combinatorics

The aim of this section is to obtain a pathwise representation for the length of the longest common and increasing subsequences of two finite strings. Throughout the paper, $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ are two infinite sequences whose coordinates take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_m\}$, a finite totally ordered alphabet of cardinality m. Recall next that LCI_n is the maximal integer $k \in \{1, \dots, n\}$, such that there exist $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$, satisfying the following two conditions:

(i) $X_{i_s} = Y_{j_s}$, for all $s = 1, 2, \dots, k$,

(ii) $X_{i_1} \leq X_{i_2} \leq \cdots \leq X_{i_k}$ and $Y_{j_1} \leq Y_{j_2} \leq \cdots \leq Y_{j_k}$.

Now that LCI_n has been formally defined, let us set some standing notation. Let $N_r(X)$, $r = 1, \dots, m$, be the number of $\boldsymbol{\alpha}_r s$ in $X_1, X_2, \dots, X_n$, i.e.,

$$N_r(X) = \#\big\{i = 1, \dots, n : X_i = \boldsymbol{\alpha}_r\big\} = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = \alpha_r\}}, \tag{2.1}$$

and similarly let $N_r(Y)$, $r = 1, \dots, m$, be the number of $\boldsymbol{\alpha}_r s$ in $Y_1, Y_2, \dots, Y_n$. Clearly,

$$\sum_{r=1}^{m} N_r(X) = \sum_{r=1}^{m} N_r(Y) = n.$$

Let us now set a convention: *Throughout the paper when there is no ambiguity or when a property is valid for both sequences $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ we often omit the symbol X or Y and, e.g., write N_r for either $N_r(X)$ or $N_r(Y)$ or, below, H for either H_X or H_Y.*

Continuing on our notational path, for each $r = 1, \dots, m$, let $N_r^{s,t}(X)$ be the number of $\boldsymbol{\alpha}_r s$ in $X_{s+1}, X_{s+2}, \dots, X_t$, i.e.,

$$N_r^{s,t}(X) = \#\big\{i = s+1, \dots, t : X_i = \boldsymbol{\alpha}_r\big\} = \sum_{i=s+1}^{t} \mathbf{1}_{\{X_i = \boldsymbol{\alpha}_r\}}, \tag{2.2}$$

with a similar definition for $N_r^{s,t}(Y)$. Again, it is trivially verified that

$$\sum_{r=1}^{m} N_r^{s,t}(X) = \sum_{r=1}^{m} N_r^{s,t}(Y) = t - s,$$

and, of course, $N_r^{0,n} = N_r$. Still continuing with our notations, let $T_r^j(X)$, $r = 1, \ldots, m$, be the location of the j^{th} $\boldsymbol{\alpha}_r$ in the *infinite* sequence $X_1, X_2, \ldots, X_n, \ldots$, with the convention that $T_r^0(X) = 0$. Then, for $j = 1, 2, \ldots$, $T_r^j(X)$ can be defined recursively via,

$$T_r^j(X) = \min\left\{s \in \mathbb{N} : s > T_r^{j-1}(X), X_s = \boldsymbol{\alpha}_r\right\} \tag{2.3}$$

where as usual $\mathbb{N} = \{0, 1, 2, \ldots\}$. Again replacing X by Y gives the corresponding notion for the sequence $Y = (Y_i)_{i \geq 1}$.

Next, let us begin our finding of a representation for LCI_n via the random variables defined to date. First, let $H_X(k_1, k_2, \ldots, k_{m-1})$ be the maximal number of $\boldsymbol{\alpha}_m s$ contained in an increasing subsequence, of $X_1 X_2 \cdots X_n$, containing k_1 $\boldsymbol{\alpha}_1 s$, k_2 $\boldsymbol{\alpha}_2 s$, $\ldots$, k_{m-1} $\boldsymbol{\alpha}_{m-1} s$ picked in that order. Replacing $X = (X_i)_{i\geq 1}$ by $Y = (Y_i)_{i \geq 1}$, it is then clear that

$$\min\Big(k_1 + \cdots + k_{m-1} + H_X(k_1, \ldots, k_{m-1}), k_1 + \cdots + k_{m-1} + H_Y(k_1, \ldots, k_{m-1})\Big), \tag{2.4}$$

is, therefore, the length of the longest common and increasing subsequence of $X_1 X_2 \cdots X_n$ and $Y_1 Y_2 \cdots Y_n$ containing exactly k_r $\boldsymbol{\alpha}_r s$, for all $r = 1, 2, \ldots, m-1$, the letters being picked in an increasing order. Hence, to find LCI_n, the function H needs to be identified and (2.4) needs to be maximized over all possible choices of $k_1, k_2, \ldots, k_{m-1}$.

Let us start with the maximizing constraints. Assume, for a while, that a single word, say, $X_1 \cdots X_n$, is considered. First, and clearly, $0 \leq k_1 \leq N_1$. Next, k_2 is the number of $\boldsymbol{\alpha}_2 s$ present in the sequence after the k_1^{th} $\boldsymbol{\alpha}_1$. Any letter $\boldsymbol{\alpha}_2$ is admissible but the ones occurring before the k_1^{th} $\boldsymbol{\alpha}_1$, attained at the location $T_1^{k_1} \wedge n$. Since there are n letters, considered so far, there are thus $N_2^{0, T_1^{k_1} \wedge n}$ inadmissible $\boldsymbol{\alpha}_2 s$ and the requirement on k_2 writes $k_2 \leq N_2 - N_2^{0, T_1^{k_1} \wedge n}$. Similarly for each $r = 3, \ldots, m-1$, k_r is the number of letters $\boldsymbol{\alpha}_r$ minus the inadmissible $\boldsymbol{\alpha}_r s$ which occur during the recuperation, of the k_1 $\boldsymbol{\alpha}_1 s$, followed by the k_2 $\boldsymbol{\alpha}_2 s$, followed by the k_3 $\boldsymbol{\alpha}_3 s$, *etc* in that order. Thus the requirement on k_r is of the form $k_r \leq N_r - \widetilde{N}_r^*$, where $\widetilde{N}_r^*$ is the number of $\boldsymbol{\alpha}_r s$ occurring before the k_i $\boldsymbol{\alpha}_i s$, $i \leq r-1$, picked in the order just described. For $r = 1, 2$, and as already shown, $\widetilde{N}_1^* = 0$ and $\widetilde{N}_2^* = N_2^{0, T_1^{k_1} \wedge n}$. Assume next that, for $r \geq 3$, $\widetilde{N}_{r-1}^*$ is well defined, then $\widetilde{N}_r^*$ is the number of $\boldsymbol{\alpha}_r s$ occurring before, in that order, the k_1 $\boldsymbol{\alpha}_1 s, \ldots,$ the k_{r-1} $\boldsymbol{\alpha}_{r-1} s$. A little moment of reflection makes it clear that the location of the k_{r-1}^{th} such $\boldsymbol{\alpha}_{r-1}$ is $T_{r-1}^{k_{r-1} + \widetilde{N}_{r-1}^*}$, from which it recursively follows that:

$$\widetilde{N}_r^* = N_r^{0, T_{r-1}^{k_{r-1} + \widetilde{N}_{r-1}^*} \wedge n}.$$

Returning to two sequences $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, the condition on k_r, $1 \leq r \leq m-1$, writes

$$0 \leq k_r \leq \Big(N_r(X) - \widetilde{N}_r^*(X)\Big) \wedge \Big(N_r(Y) - \widetilde{N}_r^*(Y)\Big).$$

From these choices of indices and (2.4),

$$\mathrm{LCI}_n = \max_{\bigcap_{i=1}^{m-1} \widetilde{\mathcal{C}}_i} \min\left(\sum_{i=1}^{m-1} k_i + H_X(k_1,\ldots,k_{m-1}), \sum_{i=1}^{m-1} k_i + H_Y(k_1,\ldots,k_{m-1})\right), \tag{2.5}$$

where, for $i = 1\ldots,m-1$,

$$\widetilde{\mathcal{C}}_i = \left\{0 \le k_i \le \left(N_i(X) - \widetilde{N}_i^*(X)\right) \bigwedge \left(N_i(Y) - \widetilde{N}_i^*(Y)\right)\right\}.$$

Next, observe that if $T_{r-1}^{k_{r-1}+\widetilde{N}_{r-1}^*} > n$, then $N_r - \widetilde{N}_r^* = 0$. Also, since the above maximum does not change under vacuous constraints, one can replace in the defining constraints, $\widetilde{N}_r^*$ by N_r^* recursively given via: $N_1^* = 0$ and for $r = 2,\ldots,m-1$,

$$N_r^* = N_r^{0,T_{r-1}^{k_{r-1}+N_{r-1}^*}}. \tag{2.6}$$

The combinatorial expression (2.5) then becomes

$$\mathrm{LCI}_n = \max_{\bigcap_{i=1}^{m-1} \mathcal{C}_i} \min\left(\sum_{i=1}^{m-1} k_i + H_X(k_1,\ldots,k_{m-1}), \sum_{i=1}^{m-1} k_i + H_Y(k_1,\ldots,k_{m-1})\right),$$

with now, for $i = 1,\ldots,m-1$,

$$\mathcal{C}_i = \left\{0 \le k_i \le \left(N_i(X) - N_i^*(X)\right) \wedge \left(N_i(Y) - N_i^*(Y)\right)\right\}, \tag{2.7}$$

and, of course,

$$\sum_{i=1}^{m} N_i(X) = \sum_{i=1}^{m} N_i(Y) = n.$$

After this identification, recall that H is the maximal number of $\boldsymbol{\alpha}_m$ after, in that order, the k_1 $\boldsymbol{\alpha}_1 s$, k_2 $\boldsymbol{\alpha}_2 s$, $\ldots$, k_{m-1} $\boldsymbol{\alpha}_{m-1} s$. Counting the $\boldsymbol{\alpha}_m s$ present between the various locations of the $\boldsymbol{\alpha}_i$, $i = 1,\ldots,m-1$, and after another moment of reflection, it is clear that

$$H = N_m - R,$$

where

$$R = \sum_{i=1}^{m-1} \sum_{j=N_i^*+1}^{N_i^*+k_i} N_m^{T_i^{j-1},T_i^j}, \tag{2.8}$$

and where the N_i^* are given by (2.6). Summarizing our results leads to:

Theorem 2.1 *Let $X = (X_i)_{i\ge1}$ and $Y = (Y_i)_{i\ge1}$ be two sequences whose coordinates take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_m\}$, a totally ordered finite alphabet of* *cardinality m. Let LCI_n be the length of the longest common and increasing subsequence of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. Then,*

$$LCI_n = \max_{\bigcap_{i=1}^{m-1} \mathcal{C}_i} \min \left(\sum_{i=1}^{m-1} k_i + N_m(X) - R(X), \sum_{i=1}^{m-1} k_i + N_m(Y) - R(Y) \right), \tag{2.9}$$

where $\mathcal{C}_i = \left\{ 0 \leq k_i \leq (N_i(X) - N_i^*(X)) \wedge (N_i(Y) - N_i^*(Y)) \right\}$, and where

$$R = \sum_{i=1}^{m-1} \sum_{j=N_i^*+1}^{N_i^*+k_i} N_m^{T_i^{j-1}, T_i^j},$$

with the various N's and T's given by (2.1), (2.2), (2.3) and (2.6).

The representation (2.9) has the great advantage of (essentially) only involving the quantities N_i, N_i^*, $i = 1, 2, \ldots, m-1$ and T_i^j, $i = 1, 2, \ldots, m-1$, $j = 1, 2, \ldots$, and N_m.

3 Probability

Let us now bring our probabilistic framework into the picture and study first the random variables $N_m^{T_i^{j-1}, T_i^j}$, $i = 1, 2, \ldots, m-1$ and $j = 1, 2, \ldots$ and then the random variables N_i^*, $i = 1, 2, \ldots, m-1$, both appearing in R in (2.8).

Lemma 3.1 *Let $(Z_n)_{n \geq 1}$ be a sequence of iid random variables with $\mathbb{P}(Z_1 = \boldsymbol{\alpha}_i) = p_i$, $i = 1, \ldots, m$. For each $i = 1, 2, \ldots, m$, let $T_i^0 = 0$, and let T_i^j, $j = 1, 2, \ldots$ be the location of the j^{th} $\boldsymbol{\alpha}_i$ in the infinite sequence $(Z_n)_{n \geq 1}$. Let $i, r \in \{1, \ldots, m\}$, with $r \neq i$. Then, for any $j = 1, 2, \ldots$, the conditional law of $N_r^{T_i^{j-1}, T_i^j}$ given (T_i^{j-1}, T_i^j), is binomial with parameters $T_i^j - T_i^{j-1} - 1$ and $p_r/(1-p_i)$, which we denote by $\mathcal{B}\big(T_i^j - T_i^{j-1} - 1, p_r/(1-p_i)\big)$. Moreover, the conditional law of $\big(N_r^{T_i^{j-1}, T_i^j}\big)_{r=1,\ldots,m, r \neq i}$ given (T_i^{j-1}, T_i^j), is multinomial with parameters $T_i^j - T_i^{j-1} - 1$ and $(p_r/(1-p_i))_{r=1,\ldots,m,r\neq i}$, which we denote by $\mathcal{M}ul\big(T_i^j - T_i^{j-1} - 1, (p_r/(1-p_i))_{r=1,\ldots,m,r\neq i}\big)$. Finally, for each $i \neq r$, the random variables $\big(N_r^{T_i^{j-1}, T_i^j}\big)_{j \geq 1}$, are independent with mean p_r/p_i and variance $(p_r/p_i)(1 + p_r/p_i)$; and, moreover, they are identically distributed in case the $(Z_n)_{n \geq 1}$, are uniformly distributed.*

Proof. Let us denote by $\mathcal{L}\big(N_r^{T_i^{j-1}, T_i^j} \big| T_i^{j-1}, T_i^j\big)$ the conditional law of $N_r^{T_i^{j-1}, T_i^j}$ given T_i^{j-1}, T_i^j. Recall, see (2.3), that T_i^{j-1} and T_i^j are the respective locations of the $(j-1)^{\text{th}}$ $\boldsymbol{\alpha}_i$ and the j^{th} $\boldsymbol{\alpha}_i$ in the infinite sequence $(Z_n)_{n \geq 1}$. Thus between $T_i^{j-1} + 1$ and T_i^j, there are $T_i^j - T_i^{j-1} - 1$ free spots and each one is equally likely contain $\boldsymbol{\alpha}_r$, $r \neq i$, with probability $p_r / (\sum_{\substack{\ell=1 \\ \ell \neq i}}^m p_\ell) = p_r/(1-p_i)$. Therefore,

$$\mathcal{L}\left(N_r^{T_i^{j-1}, T_i^j} \big| T_i^{j-1}, T_i^j \right) = \mathcal{B}\left(T_i^j - T_i^{j-1} - 1, \frac{p_r}{1 - p_i} \right). \tag{3.1}$$

Let us now compute the probability generating function of the random variables $N_r^{T_i^{j-1},T_i^j}$, $i \neq r$. First, via (3.1)

$$\begin{aligned}
\mathbb{E}\left[x^{N_r^{T_i^{j-1},T_i^j}}\right] &= \mathbb{E}\left[\mathbb{E}\left[x^{N_r^{T_i^{j-1},T_i^j}}\Big| T_i^{j-1}, T_i^j\right]\right] \\
&= \sum_{\ell=1}^{\infty}\left(1 - \frac{p_r}{1-p_i} + \frac{p_r}{1-p_i}x\right)^{\ell-1} p_i(1-p_i)^{\ell-1} \\
&= \frac{p_i}{1-(1-p_i)\left(1-\frac{p_r}{1-p_i}+\frac{p_r}{1-p_i}x\right)} \\
&= \frac{p_i}{p_i+p_r-p_r x}, \qquad (3.2)
\end{aligned}$$

since T_i^j is a negative binomial (Pascal) random variable with parameters j and p_i which we shall denote $\mathcal{BN}(j,p_i)$ in the sequel and $T_i^j - T_i^{j-1}$ is a geometric random variables with parameter p_i, which we shall denote $\mathcal{G}(p_j)$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[N_r^{T_i^{j-1},T_i^j}\right] &= \frac{p_r}{p_i}, \\
\mathrm{Var}\left(N_r^{T_i^{j-1},T_i^j}\right) &= \frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right). \qquad (3.3)
\end{aligned}$$

In the uniform case, i.e., $p_i = 1/m$, $i = 1,\ldots,m$, the $N_r^{T_i^{j-1},T_i^j}$, $i = 1,\ldots,m$, $i \neq r$, $j = 1,2,\ldots$ are clearly seen to be identically distributed, via (3.2). The multinomial part of the statement is proved in a very similar manner. The $T_i^j - T_i^{j-1} - 1$ free spots are to contain the letters $\boldsymbol{\alpha}_r$, $r \in \{1,\ldots,m\}, r \neq i$, with respective probabilities $p_r/(1-p_i)$. Therefore,

$$\mathcal{L}\left(\left(N_r^{T_i^{j-1},T_i^j}\right)_{r=1,\ldots m, r\neq i}\Big| T_i^{j-1}, T_i^j\right) = \mathcal{M}ul\left(T_i^j - T_i^{j-1} - 1, \left(\frac{p_r}{1-p_i}\right)_{r=1,\ldots,m,r\neq i}\right). \qquad (3.4)$$

Via (3.4), the probability generating function of the random vector $\left(N_r^{T_i^{j-1},T_i^j}\right)_{r=1,\ldots,m,r\neq i}$ is then given by:

$$\begin{aligned}
\mathbb{E}\left[\prod_{r=1,r\neq i}^{m} x_r^{N_r^{T_i^{j-1},T_i^j}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{\substack{r=1\\ r\neq i}}^{m} x_r^{N_r^{T_i^{j-1},T_i^j}}\Big| T_i^{j-1}, T_i^j\right]\right] \\
&= \sum_{\ell=1}^{\infty}\left(\sum_{\substack{r=1\\ r\neq i}}^{m}\frac{p_r}{1-p_i}x_r\right)^{\ell-1} p_i(1-p_i)^{\ell-1} \\
&= \frac{p_i}{1-\sum_{r=1,r\neq i}^{m} p_r x_r}. \qquad (3.5)
\end{aligned}$$

As a direct consequence of (3.5) and for $r \neq i, s \neq i$,

$$\mathrm{Cov}\left(N_r^{T_i^{j-1},T_i^j}, N_s^{T_i^{j-1},T_i^j}\right) = \frac{p_r p_s}{p_i^2}. \tag{3.6}$$

The proof of the lemma will be complete once, for each $i \neq r$, the random variables $N_r^{T_i^{j-1},T_i^j}$, $j \geq 1$, are shown to be independent. First, note that given $T_i^{j-1}, T_i^j, T_i^{k-1}, T_i^k$, the random variables $N_r^{T_i^{j-1},T_i^j} = \sum_{\ell=T_i^{j-1}+1}^{T_i^j} \mathbf{1}_{\{X_\ell=\boldsymbol{\alpha}_r\}}$ and $N_r^{T_i^{k-1},T_i^k} = \sum_{\ell=T_i^{k-1}+1}^{T_i^k} \mathbf{1}_{\{X_\ell=\boldsymbol{\alpha}_r\}}$ are independent since the intervals $[T_i^{j-1}+1, T_i^j]$ and $[T_i^{k-1}+1, T_i^k]$ are disjoint, and since the $(X_\ell)_{\ell\geq 1}$ are also independent. Moreover, recall that conditional distributions are given by (3.1), and so, for instance,

$$\begin{aligned}\mathcal{L}\left(N_r^{T_i^{j-1},T_i^j} \middle| T_i^{j-1}, T_i^j, T_i^{k-1}, T_i^k\right) &= \mathcal{L}\left(N_r^{T_i^{j-1},T_i^j} \middle| T_i^{j-1}, T_i^j\right) \\ &= \mathcal{B}\left(T_i^j - T_i^{j-1} - 1, \frac{p_r}{1-p_i}\right).\end{aligned}$$

Therefore, for any measurable functions $f, g : \mathbb{R}_+ \to \mathbb{R}_+$, and if $\mathbb{E}_{\mathcal{B}(n,p)}$ denotes the expectation with respect to a binomial $\mathcal{B}(n,p)$ distribution then

$$\begin{aligned}&\mathbb{E}\left[f\big(N_r^{T_i^{j-1},T_i^j}\big)g\big(N_r^{T_i^{k-1},T_i^k}\big)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f\big(N_r^{T_i^{j-1},T_i^j}\big)g\big(N_r^{T_i^{k-1},T_i^k}\big)\middle|T_i^{j-1}, T_i^j, T_i^{k-1}, T_i^k\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f\big(N_r^{T_i^{j-1},T_i^j}\big)\middle|T_i^{j-1}, T_i^j, T_i^{k-1}, T_i^k\right]\mathbb{E}\left[g\big(N_r^{T_i^{k-1},T_i^k}\big)\middle|T_i^{j-1}, T_i^j, T_i^{k-1}, T_i^k\right]\right] \qquad (3.7)\\ &= \mathbb{E}\left[\mathbb{E}_{\mathcal{B}\left(T_i^j-T_i^{j-1}-1,\frac{p_r}{1-p_i}\right)}[f]\ \mathbb{E}_{\mathcal{B}\left(T_i^k-T_i^{k-1}-1,\frac{p_r}{1-p_i}\right)}[g]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{\mathcal{B}\left(T_i^j-T_i^{j-1}-1,\frac{p_r}{1-p_i}\right)}[f]\right]\ \mathbb{E}\left[\mathbb{E}_{\mathcal{B}\left(T_i^k-T_i^{k-1}-1,\frac{p_r}{1-p_i}\right)}[g]\right] \qquad (3.8)\\ &= \mathbb{E}\left[f\big(N_r^{T_i^{j-1},T_i^j}\big)\right]\ \mathbb{E}\left[g\big(N_r^{T_i^{k-1},T_i^k}\big)\right],\end{aligned}$$

where the equality in (3.7) is due to the conditional independence property, while the one in (3.8) follows from that

$$\mathbb{E}_{\mathcal{B}\left(T_i^j-T_i^{j-1}-1,\frac{p_r}{1-p_i}\right)}[f] = F\big(T_i^j - T_i^{j-1}\big) \quad \text{and} \quad \mathbb{E}_{\mathcal{B}\left(T_i^k-T_i^{k-1}-1,\frac{p_r}{1-p_i}\right)}[g] = G\big(T_i^k - T_i^{k-1}\big),$$

for some functions F, G, and from the independence of $T_i^j - T_i^{j-1}$ and $T_i^k - T_i^{k-1}$. The argument can then be easily adapted to justify the mutual independence of the random variables $\big(N_r^{T_i^{j-1},T_i^j}\big)_{j\geq 1}$. □

With the help of the previous lemma and in order to prepare our first fluctuation result, it is relevant to rewrite the representation (2.9) as

$$\mathrm{LCI}_n = \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min\left\{\sum_{i=1}^{m-1} k_i + N_m(X) - G_{n,m}(X), \sum_{i=1}^{m-1} k_i + N_m(Y) - G_{n,m}(Y)\right\}, \tag{3.9}$$

where

$$G_{n,m} = \sum_{i=1}^{m-1} \sum_{j=N_i^*+1}^{N_i^*+k_i} \left(\left(\frac{N_m^{T_i^{j-1},T_i^j} - \frac{p_m}{p_i}}{\sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)n}} \right) \sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)n} + \frac{p_m}{p_i} \right), \tag{3.10}$$

and where $p_i(X) = \mathbb{P}(X_1 = \boldsymbol{\alpha}_i)$ and $p_i(Y) = \mathbb{P}(Y_1 = \boldsymbol{\alpha}_i)$, $1 \le i \le m$.

Via (3.9) and (3.10), LCI_n is now represented as a max/min over random constraints of random sums of randomly stopped independent random variables, except for the presence of $N_m(X)$ and $N_m(Y)$. Our next result also represents, up to a small error term, both $N_m(X)$ and $N_m(Y)$ via the same random variables.

Lemma 3.2 *For each $i = 1, 2, \ldots, m$, and $r \neq i$,*

$$N_r = \frac{p_r}{p_i} N_i + \sum_{j=1}^{N_i} \frac{\left(N_r^{T_i^{j-1},T_i^j} - \frac{p_r}{p_i}\right)}{\sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}} \sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n} + S_{i,r}^{(n)}, \tag{3.11}$$

where $\lim_{n\to+\infty} S_{i,r}^{(n)}/\sqrt{n} = 0$, in probability. In particular, for each $r = 1, 2, \ldots, m$,

$$N_r = np_r + \sum_{\substack{i=1\\i\neq r}}^{m} \sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}\, p_i \sum_{j=1}^{N_i} \frac{\left(N_r^{T_i^{j-1},T_i^j} - \frac{p_r}{p_i}\right)}{\sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}} + \sum_{\substack{i=1\\i\neq r}}^{m} p_i S_{i,r}^{(n)}. \tag{3.12}$$

Proof. Let us start with the proof of (3.12). Summing over $i = 1, \ldots, m$, $i \neq r$, both sides of (3.11), we get

$$\sum_{\substack{i=1\\i\neq r}}^{m} \frac{p_i}{p_r} N_r = \sum_{\substack{i=1\\i\neq r}}^{m} N_i + \sum_{\substack{i=1\\i\neq r}}^{m} \sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}\, \frac{p_i}{p_r} \left(\sum_{j=1}^{N_i} \frac{\left(N_r^{T_i^{j-1},T_i^j} - \frac{p_r}{p_i}\right)}{\sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}} \right) + \sum_{\substack{i=1\\i\neq r}}^{m} \frac{p_i}{p_r} S_{i,r}^{(n)}. \tag{3.13}$$

But, $\sum_{i=1}^m N_i = n$, and so (3.13) becomes

$$N_r = np_r + \sum_{\substack{i=1\\i\neq r}}^{m} \sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}\, p_i \left(\sum_{j=1}^{N_i} \frac{\left(N_r^{T_i^{j-1},T_i^j} - \frac{p_r}{p_i}\right)}{\sqrt{\frac{p_r}{p_i}\left(1+\frac{p_r}{p_i}\right)n}} \right) + \sum_{\substack{i=1\\i\neq r}}^{m} p_i S_{i,r}^{(n)},$$

which is precisely (3.12). Let us now prove (3.11) by identifying the random variable $S_{i,r}^{(n)}$ and show that when scaled by $\sqrt{n}$, they converge to zero in probability. Clearly, for $i = 1, \ldots, m$, $i \neq r$,

$$0 \le S_{i,r}^{(n)} := N_r - \sum_{j=1}^{N_i} N_r^{T_i^{j-1},T_i^j}.$$

In other words, $S_{i,r}^{(n)}$ is the number of $\boldsymbol{\alpha}_r$ in the interval $[T_i^*+1,n]$, where T_i^* is the location of the last $\boldsymbol{\alpha}_i$ in $[1,n]$. Therefore,

$$0 \leq S_{i,r}^{(n)} \leq n - T_i^* = n - (T_i^{N_i} \wedge n). \tag{3.14}$$

But, $\mathbb{P}(T_i^* = n-k) = p_i(1-p_i)^k$, $k = 0,1,\ldots,n-1$ and $\mathbb{P}(T_i^* = 0) = (1-p_i)^n$. Therefore, for all $\epsilon > 0$, and n large enough,

$$\mathbb{P}\left(\frac{S_{i,m}^{(n)}}{\sqrt{n}} \geq \epsilon\right) \leq \mathbb{P}(n - T_i^* \geq \epsilon\sqrt{n}) \leq \sum_{l=[\epsilon\sqrt{n}]}^{n} p_i(1-p_i)^l \leq (1-p_i)^{[\epsilon\sqrt{n}]} \underset{n\to+\infty}{\longrightarrow} 0. \tag{3.15}$$

□

Returning to the representation (3.9), the previous lemma allows us to rewrite LCI_n as:

$$\begin{aligned}\mathrm{LCI}_n = \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min\Bigg(& np_m(X) + \sum_{i=1}^{m-1} k_i - p_m(X)\sum_{i=1}^{m-1}\frac{k_i}{p_i(X)} + H_{m,n}(X) + \sum_{i=1}^{m-1} p_i(X) S_{i,m}^{(n)}(X), \\ & np_m(Y) + \sum_{i=1}^{m-1} k_i - p_m(Y)\sum_{i=1}^{m-1}\frac{k_i}{p_i(Y)} + H_{m,n}(Y) + \sum_{i=1}^{m-1} p_i(Y) S_{i,m}^{(n)}(Y)\Bigg),\end{aligned} \tag{3.16}$$

where

$$\begin{aligned}H_{m,n} = & \sum_{i=1}^{m-1}\sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)np_i}\sum_{j=1}^{N_i}\frac{\left(N_m^{T_i^{j-1},T_i^j} - \frac{p_m}{p_i}\right)}{\sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)n}} \\ & - \sum_{i=1}^{m-1}\sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)n}\sum_{j=N_i^*+1}^{N_i^*+k_i}\frac{\left(N_m^{T_i^{j-1},T_i^j} - \frac{p_m}{p_i}\right)}{\sqrt{\frac{p_m}{p_i}\left(1+\frac{p_m}{p_i}\right)n}}.\end{aligned} \tag{3.17}$$

□

We now study some properties of the random variables N_i^* which are present in both the random constraints and the random sums. The random variables N_i^* are defined recursively by (2.6) with $N_1^* = 0$. We fix $\mathbf{k} = (k_1,\ldots,k_{m-1})$ where k_i is the number of letters $\boldsymbol{\alpha}_i$ present in the common increasing subsequence. The random variables N_i^*, $i \geq 2$, depend on $\mathbf{k}$, actually $N_i^* = N_i^*(k_1,\ldots,k_{i-1})$. We write

$$N_i^* = \sum_{j=1}^{i-1} N_{i,j}^* \tag{3.18}$$

where $N_{i,j}^* = N_{i,j}^*(k_j)$ is the number of letters $\boldsymbol{\alpha}_i$ present in the step $j \le i-1$ consisting in collecting the k_j letters $\boldsymbol{\alpha}_j$, $j \le i-1$. (In the sequel, in order not to further burden the notations, we shall skip the symbols k_j, $j = 1, \dots, i-1$, in N_i^* and $N_{i,j}^*$.) The following diagram encapsulates the drawing of the letters:

$$\begin{array}{cccccccccc}
1 & T_1^{k_1} & T_2^{k_2+N_2^*} & T_3^{k_3+N_3^*} & \dots & T_{j-1}^{k_{j-1}+N_{j-1}^*} & T_j^{k_j+N_j^*} & \dots & T_{i-2}^{k_{i-2}+N_{i-2}^*} & T_{i-1}^{k_{i-1}+N_{i-1}^*} \\
\hline
k_1\ \boldsymbol{\alpha}_1 & & & & & & & & & \\
N_{2,1}^*\ \boldsymbol{\alpha}_2 & k_2\ \boldsymbol{\alpha}_2 & & & & & & & & \\
N_{3,1}^*\ \boldsymbol{\alpha}_3 & N_{3,2}^*\ \boldsymbol{\alpha}_3 & k_3\ \boldsymbol{\alpha}_3 & & & k_j\ \boldsymbol{\alpha}_j & & & & \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & & \\
 & & & & & & & & k_{i-1}\boldsymbol{\alpha}_{i-1} & \\
N_{i,1}^*\ \boldsymbol{\alpha}_i & N_{i,2}^*\ \boldsymbol{\alpha}_i & N_{i,3}^*\ \boldsymbol{\alpha}_i & \dots & & N_{i,j}^*\ \boldsymbol{\alpha}_i & \dots & & N_{i,i-1}^*\ \boldsymbol{\alpha}_i &
\end{array}$$

In Step $j \le i-1$, there are $T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*}$ letters selected but k_j letters are $\boldsymbol{\alpha}_j$, $N_{j+1,j}^*$ are $\boldsymbol{\alpha}_{j+1}$, $\dots$, $N_{i-1,j}^*$ are $\boldsymbol{\alpha}_{i-1}$, (for $j = i-1$, there are also k_j letters $\boldsymbol{\alpha}_j$ but none of the others $\boldsymbol{\alpha}_{j+1}$, *etc*).

Moreover, there are $T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j - N_{j+1,j}^* - \cdots - N_{i-1,j}^*$ possible spots ($T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j$ in case $j = i-1$) in which the probability of having a $\boldsymbol{\alpha}_i$ is $p_{i,j} := p_i/(1 - p_j - \cdots - p_{i-1})$. Therefore, conditionally on

$$\mathcal{G}_{i,j}(\mathbf{k}) = \sigma\left(N_{j+1,j}^*, \dots, N_{i-1,j}^*, T_{j-1}^{k_{j-1}+N_{j-1}^*}, T_j^{k_j+N_j^*}\right),$$

(the σ-field generated by $N_{j+1,j}^*, \dots, N_{i-1,j}^*, T_{j-1}^{k_{j-1}+N_{j-1}^*}, T_j^{k_j+N_j^*}$) it follows that

$$N_{i,j}^* \sim \mathcal{B}\Big(T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j - N_{j+1,j}^* - \cdots - N_{i-1,j}^*, p_{i,j}\Big). \tag{3.19}$$

The two forthcoming propositions respectively characterize the laws of $N_{i,j}^*$ and of N_i^*.

Proposition 3.1 *For each $i = 2, \dots, m$, the probability generating function of $N_{i,j}^*$, $1 \le j \le i-1$, is given by*

$$\mathbb{E}\Big[x^{N_{i,j}^*}\Big] = \left(\frac{p_j}{p_j + p_i - p_i x}\right)^{k_j}. \tag{3.20}$$

Therefore, $N_{i,j}^*$ is distributed as $\sum_{\ell=1}^{k_j}(G_\ell - 1)$, where $(G_\ell)_{1\le\ell\le k_j}$ are independent with geometric law $\mathcal{G}\big(p_j/(p_j+p_i)\big)$ and so,

$$\mathbb{E}[N_{i,j}^*] = \frac{p_i}{p_j}k_j \quad \text{and} \quad \operatorname{Var}(N_{i,j}^*) = \left(1 + \frac{p_i}{p_j}\right)\frac{p_i}{p_j}k_j. \tag{3.21}$$

Proof. Recall that, for $N \sim \mathcal{B}(n,p)$, $\mathbb{E}[x^N] = (1-p+px)^n$ while, for $N \sim \mathcal{G}(p)$, $\mathbb{E}[x^N] = px/(1-(1-p)x)$. Using (3.19), we then have for $N = T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j - N_{j+1,j}^* - \cdots - N_{i-1,j}^*$,

$$\begin{aligned}
\mathbb{E}\Big[x^{N_{i,j}^*}\Big] &= \mathbb{E}\Big[\mathbb{E}\big[x^{N_{i,j}^*}\big|N\big]\Big] \\
&= \mathbb{E}\Big[(1 - p_{i,j} + p_{i,j}x)^{T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j - N_{j+1,j}^* - \cdots - N_{i-1,j}^*}\Big]
\end{aligned}$$

$$= \mathbb{E}\Big[y^{U-V}\Big], \tag{3.22}$$

setting $y = (1 - p_{i,j} + p_{i,j}x)$, and

$$U := T_j^{k_j+N_j^*} - T_{j-1}^{k_{j-1}+N_{j-1}^*} - k_j \sim \mathcal{BN}(k_j, p_j) * \delta_{-k_j} \tag{3.23}$$

$$V := \sum_{r=j+1}^{i-1} N_{r,j}^* \sim \mathcal{B}\Big(U, \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j}\Big), \tag{3.24}$$

where for $j = i-1$, we also set $V = 0$. The notation $\mathcal{BN}(k,p)$ above stands for the negative binomial (Pascal) distribution with parameters k and p. The parameters of the binomial random variables V in (3.24) stem from that V counts the number of letters $\boldsymbol{\alpha}_r$, $j+1 \le r \le i-1$, between two letters $\boldsymbol{\alpha}_j$, while exactly k_j such letters are obtained, so that each $\boldsymbol{\alpha}_r$ has probability $p_r/(1-p_j)$ to appear. Hence,

$$\begin{aligned}
\mathbb{E}\Big[y^{U-V}\Big] &= \mathbb{E}\Big[\mathbb{E}\big[y^{U-V}|U\big]\Big] \\
&= \mathbb{E}\Big[y^U \mathbb{E}\big[y^{-V}|U\big]\Big] \\
&= \mathbb{E}\Big[y^U \Big(1 - \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j} + \frac{\sum_{r=j+1}^{i-1} p_r}{(1-p_j)y}\Big)^U\Big] \\
&= \mathbb{E}\Big[\Big((1 - \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j})y + \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j}\Big)^{G_1-1}\Big]^{k_j},
\end{aligned}$$

since, from (3.23), $U \sim \sum_{\ell=1}^{k_j}(G_\ell - 1)$, where the G_ℓ, $1 \le \ell \le k_j$, are iid with distribution $\mathcal{G}(p_j)$. Finally,

$$\begin{aligned}
\mathbb{E}\Big[y^{U-V}\Big] &= \left(\frac{p_j}{1 - (1-p_j)\Big(\big(1 - \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j}\big)y + \sum_{r=j+1}^{i-1} \frac{p_r}{1-p_j}\Big)}\right)^{k_j} \\
&= \left(\frac{p_j}{p_j + p_i - p_i x}\right)^{k_j},
\end{aligned}$$

since $p_{i,j} = p_i/\big(1 - \sum_{r=j}^{i-1} p_r\big)$. The expressions for the expectation and for the variance in (3.21) follow from straightforward computations. □

Recall that by convention, $N_1^* = 0$, and for $2 \le i \le m$, the following proposition gives the law of N_i^*:

Proposition 3.2 *For each $i = 2, \dots, m$, the random variables $(N_{i,j}^*)_{1 \le j \le i-1}$ are independent. Hence, the probability generating function of N_i^* is given by*

$$\mathbb{E}\Big[x^{N_i^*}\Big] = \prod_{j=1}^{i-1} \left(\frac{p_j}{p_j + p_i - p_i x}\right)^{k_j}, \tag{3.25}$$

and so,

$$\mathbb{E}[N_i^*] = \sum_{j=1}^{i-1} \frac{p_i}{p_j} k_j \qquad \textit{and} \qquad \mathrm{Var}(N_i^*) = \sum_{j=1}^{i-1} \left(1 + \frac{p_i}{p_j}\right) \frac{p_i}{p_j} k_j. \tag{3.26}$$

Proof. In view of Proposition 3.1 and of (3.18), it is enough to prove the first part of the proposition, i.e., to prove that the random variables $N_{i,j}^*$, $1 \leq j \leq i-1$, are independent. In order to simplify notations, we only show that $N_{i,1}^*$ and $N_{i,2}^*$ are independent, but the argument can easily be extended to prove the full independence property. Since the T_i^k's are stopping times, by the strong Markov property, observe that $\sigma\big(X_1, \dots, X_{T_1^{k_1}}\big) \underset{T_1^{k_1}}{\perp\!\!\!\perp} \sigma\big(X_{T_1^{k_1}+1}, \dots, X_{T_2^{k_2+N_2^*}}\big)$ where, again $\sigma(X_1, \dots, X_n)$ denotes the σ-field generated by the random variables $X_1, \dots, X_n$, while $\underset{T_1^{k_1}}{\perp\!\!\!\perp}$ stands for independence conditionally on $T_1^{k_1}$. Moreover, $T_1^{k_1}$ and $\sigma\big(X_{T_1^{k_1}+1}, \dots, X_{T_2^{k_2+N_2^*}}\big)$ are independent, and thus so are $\sigma\big(X_1, \dots, X_{T_1^{k_1}}\big)$ and $\sigma\big(X_{T_1^{k_1}+1}, \dots, X_{T_2^{k_2+N_2^*}}\big)$. The independence of $N_{i,1}^*$ and $N_{i,2}^*$ becomes clear, since $N_{i,1}^*$ is $\sigma\big(X_1, \dots, X_{T_1^{k_1}}\big)$-measurable while $N_{i,2}^*$ is $\sigma\big(X_{T_1^{k_1}+1}, \dots, X_{T_2^{k_2+N_2^*}}\big)$-measurable. The whole conclusion of the proposition then follows. $\square$

4 The Uniform Case

In this section, we specialize ours results to the case where the letters are uniformly drawn from the alphabet, i.e., $p_i(X) = p_i(Y) = 1/m$, for all $1 \leq i \leq m$. Hence, the functional LCI_n in (3.16) rewrites as

$$\mathrm{LCI}_n = \max_{\cap_{i=1}^{m-1} \mathcal{C}_i} \min \left(\frac{n}{m} + H_{m,n}(X) + \frac{1}{m} \sum_{i=1}^{m-1} S_{i,m}^{(n)}(X), \frac{n}{m} + H_{m,n}(Y) + \frac{1}{m} \sum_{i=1}^{m-1} S_{i,m}^{(n)}(Y) \right), \tag{4.1}$$

and therefore

$$\begin{aligned} \frac{\mathrm{LCI}_n - n/m}{\sqrt{2n}} \quad &= \max_{\bigcap_{i=1}^{m-1} \mathcal{C}_i} \min \left(\frac{H_{m,n}(X)}{\sqrt{2n}} + \frac{1}{m\sqrt{2n}} \sum_{i=1}^{m-1} S_{i,m}^{(n)}(X), \right. \\ &\qquad\qquad \left. \frac{H_{m,n}(Y)}{\sqrt{2n}} + \frac{1}{m\sqrt{2n}} \sum_{i=1}^{m-1} S_{i,m}^{(n)}(Y) \right). \end{aligned} \tag{4.2}$$

The following simple inequality, a version of which is already present in [HLM], will be of multiple use:

Lemma 4.1 *Let a_k, b_k, c_k, d_k, $1 \leq k \leq K$, be reals. Then,*

$$\left| \max_{k=1,\dots,K} (a_k \wedge b_k) - \max_{k=1,\dots,K} \big((a_k + c_k) \wedge (b_k + d_k)\big) \right| \leq \max_{k=1,\dots,K} \Big(|c_k| \vee |d_k|\Big). \tag{4.3}$$

Proof First,

$$\left|\max_{k=1,\dots,K}\big(a_k\wedge b_k\big)-\max_{k=1,\dots,K}\big((a_k+c_k)\wedge(b_k+d_k)\big)\right|$$
$$\le \max_{k=1,\dots,K}\left|\big(a_k\wedge b_k\big)-\big((a_k+c_k)\wedge(b_k+d_k)\big)\right|.$$

Next, the result will follows from the elementary inequality

$$(a\wedge b)-(a+c)\wedge(b+d)\le |c|\vee|d|, \tag{4.4}$$

which is valid for all $a,b,c,d\in\mathbb{R}$. Indeed, set $D=(a\wedge b)-(a+c)\wedge(b+d)$ and assume (without loss of generality) that $a\le b$. If $a+c\le b+d$, then $D=a-(a+c)=-c\le|c|$. If $b+d\le a+c$, then $D=a-b-d$ and so whenever $a\le b+d$, (4.4) is immediate, while if $a\ge b+d$, then $D=a-b-d\le -d=|d|$ since $a-b\le 0$ and $-d\ge b-a\ge 0$. □

The previous lemma entails

$$\left|\max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i}\min\left(\frac{H_{m,n}(X)}{\sqrt{2n}}+\frac{1}{m\sqrt{2n}}\sum_{i=1}^{m-1}S_{i,m}^{(n)}(X),\frac{H_{m,n}(Y)}{\sqrt{2n}}+\frac{1}{m\sqrt{2n}}\sum_{i=1}^{m-1}S_{i,m}^{(n)}(Y)\right)\right.$$
$$\left.-\max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i}\min\left(\frac{H_{m,n}(X)}{\sqrt{2n}},\frac{H_{m,n}(Y)}{\sqrt{2n}}\right)\right|$$
$$\le\frac{1}{m\sqrt{2n}}\left(\left|\sum_{i=1}^{m-1}S_{i,m}^{(n)}(X)\right|\vee\left|\sum_{i=1}^{m-1}S_{i,m}^{(n)}(Y)\right|\right).$$

But, from Lemma 3.2, as $n\to+\infty$, both $S_{i,m}^{(n)}(X)/\sqrt{n}\xrightarrow{\mathbb{P}}0$ and $S_{i,m}^{(n)}(Y)/\sqrt{n}\xrightarrow{\mathbb{P}}0$, for all $1\le i\le m-1$ (see (3.15)). Therefore, the fluctuations of LCI_n expressed in (4.2) are the same as that of

$$\max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i}\min\left(\frac{H_{m,n}(X)}{\sqrt{2n}},\frac{H_{m,n}(Y)}{\sqrt{2n}}\right).$$

For uniform draws, the functional $H_{m,n}$ in (3.17) rewrites as

$$\begin{aligned}
H_{m,n} &= \sum_{i=1}^{m-1}\sqrt{2n}\frac{1}{m}\sum_{j=1}^{N_i}\frac{N_m^{T_i^{j-1},T_i^j}-1}{\sqrt{2n}}-\sum_{i=1}^{m-1}\sqrt{2n}\sum_{j=N_i^*+1}^{N_i^*+k_i}\frac{N_m^{T_i^{j-1},T_i^j}-1}{\sqrt{2n}}\\
&= \sqrt{2n}\left(\frac{1}{m}\sum_{i=1}^{m-1}B_n^i\left(\frac{N_i}{n}\right)-\sum_{i=1}^{m-1}\left(B_n^i\left(\frac{N_i^*+k_i}{n}\right)-B_n^i\left(\frac{N_i^*}{n}\right)\right)\right)
\end{aligned}$$

where B_n^i is the Brownian approximation defined from the random variables $N_m^{T_i^{j-1},T_i^j}$, $j\ge1$, which are iid, by Lemma 3.1, centered and scaled to have variance one, i.e., B_n is the polygonal process on $[0,1]$ defined by linear interpolation between the values

$$B_n^i\left(\frac{k}{n}\right)=\sum_{j=1}^{k}\frac{Z_j^{(i)}}{\sqrt{n}} \tag{4.5}$$

where

$$Z_j^{(i)} = \frac{N_m^{T_i^{j-1},T_i^j} - 1}{\sqrt{2}}. \tag{4.6}$$

Next, we present some heuristic arguments which provide the limiting behavior of

$$\begin{aligned}\max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min \Bigg(& \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{N_i(X)}{n}\right) - \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right)\right), \\ & \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{N_i(Y)}{n}\right) - \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{N_i^*(Y)+k_i}{n}\right) - B_n^{i,Y}\left(\frac{N_i^*(Y)}{n}\right)\right)\Bigg), \end{aligned} \tag{4.7}$$

knowing that, by Donsker theorem, $(B_n^1, \dots, B_n^{m-1}) \overset{(C_0([0,1]))^{m-1}}{\Longrightarrow} (B^1, \dots, B^{m-1})$, $n \to +\infty$, where $(B^1, \dots, B^{m-1})$ is a drift-less, $(m-1)$-dimensional, correlated Brownian motion on $[0,1]$, which is also zero at the origin. The correlation structure of this multivariate Brownian motion is given by that of the $Z_j^{(i)}$, $1 \le i \le m-1$, which in turn is given by Lemma 3.1. (Above, $\overset{(C_0([0,1]))^{m-1}}{\Longrightarrow}$ stands for the convergence in law in the product space of continuous function on $[0,1]$ vanishing at the origin.)

Heuristics

Roughly speaking, there are three limits to handle in (4.7):

1. The limit of the constraints in the maximum, $\bigcap_{i=1}^{m-1}\mathcal{C}_i$;
2. The limit of the linear terms: $\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{N_i(X)}{n}\right)$;
3. The limit of the increments: $\sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right)\right)$;

and, similarly, for X replaced by Y. Below, the symbol $\rightsquigarrow$ indicates an heuristic replacement or an heuristic limit, as $n \to +\infty$.

First Limit (to be treated last, in Section 4.3): Since $\mathcal{C}_i = \{\boldsymbol{k} = (k_1, \dots, k_{m-1}) : 0 \le k_i \le \min\left(N_i(X) - N_i^*(X), N_i(Y) - N_i^*(Y)\right)\}$, (and, again, with vacuous constraints in case either $N_i^*(X) > n$ or $N_i^*(Y) > n$) and from the concentration property of the N_i^*, we expect (with again $k_0 = 0$, and $t_0 = 0$, below):

$$\begin{aligned}\mathcal{C}_i \quad &\rightsquigarrow \quad \left\{\boldsymbol{k} = (k_1, \dots, k_{m-1}) : 0 \le k_i \le \left(\mathbb{E}[N_i(X)] - \sum_{j=1}^{i-1} k_j\right) \wedge \left(\mathbb{E}[N_i(Y)] - \sum_{j=1}^{i-1} k_j\right)\right\} \\ &= \quad \left\{\boldsymbol{k} = (k_1, \dots, k_{m-1}) : \frac{1}{n}\sum_{j=1}^{i-1} k_j \le \frac{1}{n}\sum_{j=1}^{i} k_j \le \frac{\mathbb{E}[N_i]}{n}, i = 1, \dots, m-1\right\}.\end{aligned}$$

Hence,

$$\bigcap_{i=1}^{m-1} \mathcal{C}_i \rightsquigarrow \mathcal{V}\left(\frac{1}{m}, \ldots, \frac{1}{m}\right),$$

where $\mathcal{V}(p_1, \ldots, p_{m-1}) = \{\boldsymbol{t} = (t_1, \ldots, t_{m-1}) : t_i \geq 0, i = 1, \ldots, m-1, t_1 \leq p_1, t_1 + t_2 \leq p_2, \ldots, t_1 + \cdots + t_{m-1} \leq p_{m-1}\}$.

Second Limit (see Section 4.1): For each $i = 1, \ldots, m-1$, the random variables N_i are concentrated around their respective mean $\mathbb{E}[N_i]$ $(= 1/m)$, and so

$$\frac{N_i}{n} \rightsquigarrow \mathbb{E}[N_i] \quad \text{and} \quad \sum_{i=1}^{m-1} B_n^i\left(\frac{N_i}{n}\right) \rightsquigarrow \sum_{i=1}^{m-1} B^i\big(\mathbb{E}[N_i]\big) = \sum_{i=1}^{m-1} B^i\Big(\frac{1}{m}\Big),$$

where the limit $B_n^i \xRightarrow{C_0([0,1])} B^i$ is taken simultaneously.

Third Limit (see Section 4.2): For each $i = 1, \ldots, m-1$, the random variables N_i^* are also concentrated around their mean $\mathbb{E}[N_i^*] = \sum_{j=1}^{i-1} k_j$, and so $N_i^* \rightsquigarrow \sum_{j=1}^{i-1} k_j$. Therefore,

$$\begin{aligned} B_n^{i,X}\left(\frac{N_i^*(X) + k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right) &\rightsquigarrow B_n^{i,X}\left(\sum_{j=1}^{i} \frac{k_j}{n}\right) - B_n^{i,X}\left(\sum_{j=1}^{i-1} \frac{k_j}{n}\right) \\ &\rightsquigarrow B^{i,X}\left(\sum_{j=1}^{i} t_j\right) - B^{i,X}\left(\sum_{j=1}^{i-1} t_j\right), \end{aligned}$$

and similarly for X replaced by Y. Hence,

$$\begin{aligned} \frac{\mathrm{LCI}_n - n/m}{\sqrt{2n}} \rightsquigarrow \max_{\mathcal{V}(1/m,\ldots,1/m)} \min\Bigg(&\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}\Big(\frac{1}{m}\Big) - \sum_{i=1}^{m-1}\Bigg(B^{i,X}\Bigg(\sum_{j=1}^{i} t_j\Bigg) - B^{i,X}\Bigg(\sum_{j=1}^{i-1} t_j\Bigg)\Bigg), \\ &\frac{1}{m}\sum_{i=1}^{m-1} B^{i,Y}\left(\frac{1}{m}\right) - \sum_{i=1}^{m-1}\left(B^{i,Y}\left(\sum_{j=1}^{i} t_j\right) - B^{i,Y}\left(\sum_{j=1}^{i-1} t_j\right)\right)\Bigg) \\ \stackrel{\mathcal{L}}{=} \frac{1}{\sqrt{m}} \max_{0=u_0 \leq u_1 \leq \cdots \leq u_{m-1} \leq 1} \min\Bigg(&\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}(1) - \sum_{i=1}^{m-1}\big(B^{i,X}(u_i) - B^{i,X}(u_{i-1})\big), \\ &\frac{1}{m}\sum_{i=1}^{m-1} B^{i,Y}(1) - \sum_{i=1}^{m-1}\big(B^{i,Y}(u_i) - B^{i,Y}(u_{i-1})\big)\Bigg), \end{aligned}$$

by Brownian scaling and the reparametrization $\sum_{j=1}^{i} t_j = u_i/m$, $i = 1, \ldots m-1$, $u_0 = t_0 = 0$. In other words,

$$\frac{\mathrm{LCI}_n - n/m}{\sqrt{2n/m}} \rightsquigarrow \max_{0=u_0 \leq u_1 \leq \cdots \leq u_{m-1} \leq 1} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}(1) - \sum_{i=1}^{m-1}\big(B^{i,X}(u_i) - B^{i,X}(u_{i-1})\big),\right.$$

$$\frac{1}{m}\sum_{i=1}^{m-1} B^{i,Y}(1) - \sum_{i=1}^{m-1}\left(B^{i,Y}(u_i) - B^{i,Y}(u_{i-1})\right)\Bigg).$$

Finally, a linear transformation and Brownian properties allow to transform the parameter space into the Weyl chamber

$$\mathcal{W}_m(1) := \big\{\mathbf{u} = (u_0, u_1, \dots, u_{m-1}, u_m) : 0 = u_0 \le u_1 \le \cdots \le u_{m-1} \le u_m = 1\big\},$$

and to replace the $(m-1)$-dimensional correlated Brownian motion B^X (resp. B^Y), by an m-dimensional standard one B_1 (resp. B_2). Combining these facts, the expression on the right-hand side above, becomes equal, in law, to:

$$\max_{(u_i)_{0\le i\le m}\in\mathcal{W}_m(1)} \min\Bigg(-\frac{1}{m}\sum_{i=1}^{m} B_1^{(i)}(1) + \sum_{i=1}^{m}\Big(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\Big), \\ -\frac{1}{m}\sum_{i=1}^{m} B_2^{(i)}(1) + \sum_{i=1}^{m}\Big(B_2^{(i)}(u_i) - B_2^{(i)}(u_{i-1})\Big)\Bigg),$$

which is the final form of our result, Theorem 1.1. In the sequel, we make precise the previous heuristic arguments.

4.1 The Linear Terms

Set

$$R(X) = \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^* + k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*}{n}\right)\right),$$

so that with the help of (4.7), (4.2) rewrites

$$\frac{\mathrm{LCI}_n - n/m}{\sqrt{2n}} = \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{N_i(X)}{n}\right) - R(X), \right. \\ \left. \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{N_i(Y)}{n}\right) - R(Y)\right) + o_{\mathbb{P}}(1), \quad (4.8)$$

where, throughout, $o_{\mathbb{P}}(1)$ indicates a term, which might be different from an expression to another, converging to zero, in probability, as n converges to infinity.

Next, by Lemma 4.1,

$$\Bigg|\max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\Big(\frac{N_i(X)}{n}\Big) - R(X), \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\Big(\frac{N_i(Y)}{n}\Big) - R(Y)\right)$$

$$
\begin{aligned}
&\qquad - \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{\mathbb{E}[N_i(X)]}{n}\right) - R(X), \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{\mathbb{E}[N_i(Y)]}{n}\right) - R(Y)\right)\Bigg| \\
&\le \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg|\min\left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{N_i(X)}{n}\right) - R(X), \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{N_i(Y)}{n}\right) - R(Y)\right) \\
&\qquad\qquad - \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{\mathbb{E}[N_i(X)]}{n}\right) - R(X), \frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{\mathbb{E}[N_i(Y)]}{n}\right) - R(Y)\right)\Bigg| \\
&\le \max_{\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg(\max\Bigg(\frac{1}{m}\left|\sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i(X)}{n}\right) - B_n^{i,X}\left(\frac{\mathbb{E}[N_i(X)]}{n}\right)\right)\right|, \\
&\qquad\qquad \frac{1}{m}\left|\sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{N_i(Y)}{n}\right) - B_n^{i,Y}\left(\frac{\mathbb{E}[N_i(Y)]}{n}\right)\right)\right|\Bigg)\Bigg). \qquad (4.9)
\end{aligned}
$$

We now wish to show that the right-hand side of (4.9) converges to zero, in probability. First note that for each $1 \le i \le m-1$, $\mathcal{C}_i \subset \big\{\boldsymbol{k} = (k_1, \ldots, k_{m-1}) : 0 \le k_i \le \min(N_i(X), N_i(Y))\big\} \subset \big\{\boldsymbol{k} = (k_1, \ldots, k_{m-1}) : 0 \le k_i \le n\big\}$, see (2.7). But, $B_n^i(N_i/n) - B_n^i(\mathbb{E}[N_i]/n)$, where we have dropped X and Y, does not depend on $\boldsymbol{k}$. Therefore, the maximum can be skipped and the problem reduces to showing that, for all $1 \le i \le m-1$:

$$
\left|B_n^i\left(\frac{N_i}{n}\right) - B_n^i\left(\frac{\mathbb{E}[N_i]}{n}\right)\right| \xrightarrow{\mathbb{P}} 0, \qquad (4.10)
$$

as $n \to +\infty$. This follows from the forthcoming lemma applied, for each $i = 1, \ldots, m-1$, to the random variables $Z_j^{(i)} = \big(N_m^{T_i^{j-1}, T_i^j} - 1\big)/\sqrt{2}$, present in both (4.5) and (4.6) and which, by Lemma 3.1, are iid with mean zero and variance one. Note that the lemma below can indeed be brought into play since Hoeffding's inequality, applied to the random variables N_i, ensures that

$$
\lim_{n\to+\infty} \mathbb{P}\big(|N_i - \mathbb{E}[N_i]| \ge \sqrt{n}\ln n\big) \le \lim_{n\to+\infty} \frac{2}{n^2} = 0. \qquad (4.11)
$$

Lemma 4.2 *Let $(Z_j)_{j\ge1}$ be iid centered random variables with unit variance, and for each $n \in \mathbb{N}$, let $N^{(n)}$ be an $\mathbb{N}$-valued random variable such that $\lim_{n\to+\infty} \mathbb{P}\big(|N^{(n)} - \mathbb{E}[N^{(n)}]| \ge \sqrt{n}\ln n\big) = 0$. Then,*

$$
\sum_{j\in[N^{(n)}, \mathbb{E}[N^{(n)}]]} \frac{Z_j}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0,
$$

where $[N^{(n)}, \mathbb{E}[N^{(n)}]]$ is short for $[\min(N^{(n)}, \mathbb{E}[N^{(n)}]), \max(N^{(n)}, \mathbb{E}[N^{(n)}])]$.

Proof. Let $C_n = \big\{\, \big|N^{(n)} - \mathbb{E}[N^{(n)}]\big| < \sqrt{n}\ln n\big\}$, and for $\varepsilon > 0$, let

$$
A_n(\varepsilon) = \Big\{\Big|\textstyle\sum_{j\in[N^{(n)}, \mathbb{E}[N^{(n)}]]} \frac{Z_j}{\sqrt{n}}\Big| \ge \varepsilon\Big\}.
$$

Since $\mathbb{P}\big(A_n(\varepsilon)\big) \le \mathbb{P}\big(A_n(\varepsilon) \cap C_n\big) + \mathbb{P}(C_n^c)$, and since $\lim_{n\to\infty} \mathbb{P}(C_n^c) = 0$. and it is enough to show $\lim_{n\to+\infty} \mathbb{P}\big(A_n(\varepsilon) \cap C_n\big) = 0$. But, by Kolmogorov's maximal inequality,

$$\begin{aligned}
\mathbb{P}\big(A_n(\varepsilon) \cap C_n\big) &\le \mathbb{P}\left(\max_{|k-\mathbb{E}[N^{(n)}]|<\sqrt{n}\ln n} \Big| \sum_{j\in[k,\mathbb{E}[N^{(n)}]]} \frac{Z_j}{\sqrt{n}} \Big| \ge \varepsilon \right) \\
&\le \frac{\sqrt{n}\ln n \operatorname{Var}(Z_1)}{\varepsilon^2 n} \to 0, \quad n \to +\infty.
\end{aligned}$$

□

At this stage, (4.10) is proved and therefore,

$$\begin{aligned}
\frac{\mathrm{LCI}_n - n/m}{\sqrt{2n}} = \max_{\bigcap_{i=1}^{m-1} \mathcal{C}_i} \min &\left(\frac{1}{m} \sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{\mathbb{E}[N_i(X)]}{n}\right) - R(X), \right. \\
&\left. \frac{1}{m} \sum_{i=1}^{m-1} B_n^{i,Y}\left(\frac{\mathbb{E}[N_i(Y)]}{n}\right) - R(Y) \right) + o_{\mathbb{P}}(1),
\end{aligned} \tag{4.12}$$

finishing the first part of the proof of Theorem 1.1. Indeed, $(N_1, \dots, N_m)$ is multinomial with parameters n and $(p_1, \dots, p_m)$. So, for uniform draws, $\mathbb{E}[N_i(X)] = \mathbb{E}[N_i(Y)] = np_i = n/m$. Then, by the multivariate CLT (or more precisely, the multivariate Donsker theorem and scaling), $\sum_{i=1}^{m-1} B_n^{i,X}\left(\frac{\mathbb{E}[N_i(X)]}{n}\right)/m$ converges in distribution, as $n \to +\infty$, to $\sum_{i=1}^{m-1} B_1^{i,X}(1)/(m\sqrt{m})$ where, in view of (3.6), $\big(B_1^{1,X}(t), \dots, B_1^{m-1,X}(t)\big)_{0\le t\le 1}$ is a $(m-1)$-dimensional Brownian motion with covariance matrix $t\Sigma = t(\sigma_{i,j})_{1\le i,j\le m-1}$, with $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 1/2$, $i, j = 1, \dots, m-1$. A similar result also holds replacing X by Y.

The multivariate Donsker theorem mentioned above, easily derives from the univariate one and from the multivariate CLT as follows. First, the latter gives the convergence of the finite-dimensional distributions of $(B_n^{1,X}(t), \dots, B_n^{m-1,X}(t))_{0\le t\le 1}$ with a covariance structure given by that of the $Z_1^{(i)}$, $1 \le i \le m-1$, in (4.6), see (3.6). Second, the tightness of $\big(B_n^{1,X}(t), \dots, B_n^{m-1,X}(t)\big)_{0\le t\le 1}$ is obtained from that of its coordinates: since $B_n^{i,X}$ is tight for each $1 \le i \le m-1$ by the univariate Donsker theorem, for all $\varepsilon > 0$, there is a compact K_i of $C_0([0,1])$, the usual space of continuous functions on $[0,1]$ vanishing at the origin, such that $\sup_{n\ge 1} \mathbb{P}\big(B_n^{i,X} \notin K_i\big) < \varepsilon$ and we have

$$\sup_{n\ge 1} \mathbb{P}\Big(\big(B_n^{1,X}, \dots, B_n^{m-1,X}\big) \notin K_1 \times \cdots \times K_{m-1} \Big) \le \sup_{n\ge 1} \sum_{i=1}^{m-1} \mathbb{P}\big(B_n^{i,X} \notin K_i\big) < (m-1)\varepsilon$$

with $K_1 \times \cdots \times K_{m-1}$ compact of $(C_0([0,1]))^{m-1}$ so that $\big(B_n^{1,X}, \dots, B_n^{m-1,X}\big)$ is tight. □

4.2 The Increments

In this section, we compare the maximum of two different quantities over the same set of constraints in order to simplify the quantities to be maximized (before simplifying the constraints $\bigcap_{i=1}^{m-1} \mathcal{C}_i$, themselves, in the next section). The quantities to compare are:

$$\max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg\{ \left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}(p_i(X)) - \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right)\right)\right) \bigwedge \\ \left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}(p_i(Y)) - \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{N_i^*(Y)+k_i}{n}\right) - B_n^{i,Y}\left(\frac{N_i^*(Y)}{n}\right)\right)\right)\Bigg\}, \tag{4.13}$$

and

$$\max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg\{ \left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,X}(p_i(X)) - \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\right) \bigwedge \\ \left(\frac{1}{m}\sum_{i=1}^{m-1} B_n^{i,Y}(p_i(Y)) - \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\right)\Bigg\}. \tag{4.14}$$

Using (4.3) in Lemma 4.1, their absolute difference is upper-bounded by

$$\begin{aligned}
&\max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg\{ \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right)\right) \\
&\qquad\qquad - \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\Bigg| \\
&\qquad \bigvee \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{N_i^*(Y)+k_i}{n}\right) - B_n^{i,Y}\left(\frac{N_i^*(Y)}{n}\right)\right) \\
&\qquad\qquad - \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\Bigg| \Bigg\} \\
&\le \max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg\{ \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{(N_i^*(X)+k_i)}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right)\right)\Bigg| \\
&\qquad\qquad \bigvee \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{(N_i^*(Y)+k_i)}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right)\right)\Bigg| \Bigg\} \\
&+ \max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i} \Bigg\{ \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\Bigg| \\
&\qquad\qquad \bigvee \Bigg| \sum_{i=1}^{m-1}\left(B_n^{i,Y}\left(\frac{N_i^*(Y)}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right)\Bigg| \Bigg\}.
\end{aligned}$$

Recall that $N_1^*(X) = N_1^*(Y) = 0$. Hence, for $i = 1$,

$$B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right) = B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right) = 0,$$

with the same property for functionals relative to Y. Therefore, we are left with investigating terms of the form

$$\max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i}\left\{\left|B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right)\right| \right.$$
$$\left. \bigvee \left|B_n^{i,Y}\left(\frac{N_i^*(Y)+k_i}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right)\right|\right\}, \tag{4.15}$$

and

$$\max_{\boldsymbol{k}\in\bigcap_{i=1}^{m-1}\mathcal{C}_i}\left\{\left|B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right| \bigvee \left|B_n^{i,Y}\left(\frac{N_i^*(Y)}{n}\right) - B_n^{i,Y}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right|\right\}, \tag{4.16}$$

for $2 \le i \le m-1$. Above, all the quantities considered only depend on a single sequence, say X or Y, except for the constraints in $\mathcal{C}$ which depend on both X and Y. However, $\mathcal{C}_i \subset \mathcal{C}_i^*(X) := \left\{\boldsymbol{k} = (k_1, \ldots, k_{m-1}) : 0 \le k_i \le N_i(X) - N_i^*(X)\right\}$ (resp. $\mathcal{C}_i \subset \mathcal{C}_i^*(Y)$)) and so upper-bounding, in (4.15) and (4.16), the inner maxima by sums and the maxima over $\mathcal{C}$ by maxima over $\mathcal{C}^*(X) := \bigcap_{i=1}^{m-1}\mathcal{C}_i^*(X)$ (resp. $\mathcal{C}^*(Y)$) we are left with investigating, for $2 \le i \le m-1$, the convergence in probability of terms of the form

$$\max_{\boldsymbol{k}\in\mathcal{C}^*(X)}\left\{\left|B_n^{i,X}\left(\frac{N_i^*(X)+k_i}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i} k_j}{n}\right)\right|\right\}, \tag{4.17}$$

and

$$\max_{\boldsymbol{k}\in\mathcal{C}^*(X)}\left\{\left|B_n^{i,X}\left(\frac{N_i^*(X)}{n}\right) - B_n^{i,X}\left(\frac{\sum_{j=1}^{i-1} k_j}{n}\right)\right|\right\}, \tag{4.18}$$

and, similarly with X replaced by Y. Omitting the reference to either X or Y, the terms to control are, from (4.5) and for each, $2 \le i \le m-1$, of the form:

$$\max_{\boldsymbol{k}\in\mathcal{C}^*}\left|\sum_{j=k_1+\cdots+k_i+1}^{N_i^*+k_i}\frac{Z_j^{(i)}}{\sqrt{n}}\right|, \tag{4.19}$$

and

$$\max_{\boldsymbol{k}\in\mathcal{C}^*}\left|\sum_{j=k_1+\cdots+k_{i-1}+1}^{N_i^*}\frac{Z_j^{(i)}}{\sqrt{n}}\right|, \tag{4.20}$$

where the $Z_j^{(i)}$, $j \geq 1$, are defined in (4.6) and where $\mathcal{C}^* = \bigcap_{i=1}^{m-1} \mathcal{C}_i^*$. Since (4.20) is similar to, but easier to tackle than (4.19), we only deal with (4.19). Again, as in Section 4.1, let $C_n^i = \left\{ |N_i - \mathbb{E}[N_i]| \leq \sqrt{n} \ln n \right\}$ for $i = 1, 2, \ldots, m-1$, and, thus, for $\varepsilon > 0$,

$$\begin{aligned}
& \mathbb{P}\left(\max_{\boldsymbol{k} \in \mathcal{C}^*} \left| \sum_{j=k_1+\cdots+k_i+1}^{N_i^*+k_i} \frac{Z_j^{(i)}}{\sqrt{n}} \right| \geq \varepsilon \right) \\
\leq \; & \mathbb{P}\left(\max_{\boldsymbol{k} \in \mathcal{C}^*} \left| \sum_{j=k_1+\cdots+k_i+1}^{N_i^*+k_i} Z_j^{(i)} \right| \geq \varepsilon\sqrt{n}, \bigcap_{i=1}^{m-1} C_n^i \right) + \sum_{i=1}^{m-1} \mathbb{P}\big((C_n^i)^c\big) \\
\leq \; & \mathbb{P}\left(\max_{\boldsymbol{k} \in \mathcal{C}_{n,i}^{**}} \left| \sum_{j=\ell_i+1}^{\ell_i+N_i^*-(k_1+\cdots+k_{i-1})} Z_j^{(i)} \right| \geq \varepsilon\sqrt{n} \right) + \sum_{i=1}^{m-1} \mathbb{P}\big((C_n^i)^c\big),
\end{aligned} \tag{4.21}$$

where

$$\begin{aligned}
\mathcal{C}_{n,i}^{**} = & \big\{ k_1 \leq \mathbb{E}[N_1] + \sqrt{n}\ln n, k_2 \leq \mathbb{E}[N_2] + \sqrt{n}\ln n, \ldots, k_{i-1} \leq \mathbb{E}[N_{i-1}] + \sqrt{n}\ln n, \\
& \quad k_1 + \cdots + k_{i-1} \leq \ell_i := k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - (k_1 + \cdots + k_{i-1}))\big\} \\
& = \bigcap_{j=1}^{i-1} \big\{ k_j \leq \mathbb{E}[N_j] + \sqrt{n}\ln n \big\} \\
& \quad \cap \big\{ \mathbb{E}[N_i^*] \leq \ell_i = k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - \mathbb{E}[N_i^*]) \big\} \\
& := \mathcal{C}_n^{i-1} \cap \big\{ \mathbb{E}[N_i^*] \leq \ell_i = k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - \mathbb{E}[N_i^*]) \big\},
\end{aligned} \tag{4.22}$$

since $N_i^* = N_i^*(k_1, \ldots, k_{i-1})$ is such that $\mathbb{E}[N_i^*] = k_1 + \cdots + k_{i-1}$. Now, in view of (4.11), it is enough to show the convergence to zero of the first term on the right-hand side of (4.21). To do so, let

$$D_n^i(k_1, \ldots, k_{i-1}) = \big\{ |N_i^*(k_1, \ldots, k_{i-1}) - \mathbb{E}[N_i^*(k_1, \ldots, k_{i-1})]| \leq x_n \big\},$$

with

$$x_n = n^\beta, \quad \beta \in (1/2, 1), \tag{4.23}$$

and let

$$\mathcal{D}_n^i = \bigcap_{(k_1, \ldots, k_{i-1}) \in \mathcal{C}_n^{i-1}} D_n^i(k_1, \ldots, k_{i-1}). \tag{4.24}$$

Our next goal is to show that asymptotically, $\mathcal{D}_n^i$ has full probability. To do so, we first need some technical results.

Lemma 4.3 *For $x \in [-n, +\infty)$, let*

$$F_n(x) = \frac{(x+2n)^{x+2n}}{(2x+2n)^{x+n}(2n)^n}.$$

Then, for some constants $c, C \in (0, +\infty)$,

$$F_n(x) \leq C \exp\left(-cn \min\left(\frac{|x|}{n}, \frac{x^2}{n^2}\right)\right). \tag{4.25}$$

Proof. Consider three cases: $|x| \ll n$, $x \gg n$ and $x \approx n$, i.e., $c_1 n \leq x \leq c_2 n$ for two finite constants c_1 and c_2, and expand $F_n(x)$ accordingly. First, let $|x| \ll n$: then,

$$\begin{aligned}
F_n(x) &= \frac{(2n)^{x+2n}}{(2n)^{x+n}(2n)^n} \frac{(1+\frac{x}{2n})^{x+2n}}{(1+\frac{x}{n})^{x+n}} \\
&= \exp\left((x+2n)\ln\left(1+\frac{x}{2n}\right) - (x+n)\ln\left(1+\frac{x}{n}\right)\right) \\
&= \exp\left((x+2n)\left(\frac{x}{2n} - \frac{x^2}{8n^2} + o\left(\frac{x^2}{n^2}\right)\right) - (x+n)\left(\frac{x}{n} - \frac{x^2}{2n^2} + o\left(\frac{x^2}{n^2}\right)\right)\right) \\
&= \exp\left(-\frac{x^2}{4n} + \frac{3x^3}{8n^2} + o\left(\frac{x^3}{n^2}\right) + o\left(\frac{x^2}{n}\right)\right) \\
&= \exp\left(-\frac{x^2}{4n} + o\left(\frac{x^2}{n}\right)\right),
\end{aligned}$$

which yields (4.25) in case $|x| \ll n$. Next, let $x \gg n$: then,

$$\begin{aligned}
F_n(x) &= \frac{(x+2n)^{x+2n}}{(2x+2n)^{x+n}(2n)^n} = \frac{x^n}{(4n)^n 2^x} \frac{(1+\frac{2n}{x})^{x+2n}}{(1+\frac{n}{x})^{x+n}} \\
&= \frac{x^n}{(4n)^n 2^x} \exp\left((x+2n)\ln\left(1+\frac{2n}{x}\right) - (x+n)\ln\left(1+\frac{n}{x}\right)\right) \\
&= \frac{x^n}{(4n)^n 2^x} \exp\left((x+2n)\left(\frac{2n}{x} - \frac{2n^2}{x^2} + o\left(\frac{n^2}{x^2}\right)\right) - (x+n)\left(\frac{n}{x} - \frac{n^2}{2x^2} + o\left(\frac{n^2}{x^2}\right)\right)\right) \\
&= \frac{x^n}{(4n)^n 2^x} \exp\left(n + \frac{3n^2}{2x} - \frac{7n^3}{2x^2} + o\left(\frac{n^2}{x}\right)\right) \\
&= \exp\left(n + \frac{3n^2}{2x} + n\ln\left(\frac{x}{4n}\right) - x\ln 2 + o\left(\frac{n^2}{x}\right)\right).
\end{aligned} \tag{4.26}$$

Since $x \gg n$, the larger order in the exponential (4.26) is $x \ln 2$ and, this recover a bound of the form (4.25) in this case. Finally, consider the case $x \approx n$, say $x = \alpha n$ with $\alpha > -1$. Then,

$$F_n(x) = \frac{\big((\alpha+2)n\big)^{(\alpha+2)n}}{\big((2\alpha+2)n\big)^{(\alpha+1)n}(2n)^n} = \exp\big(-c(\alpha)n\big),$$

which is again of the form (4.25), since $c(\alpha) = \ln\big(2(2\alpha+2)^{\alpha+1}/(\alpha+2)^{\alpha+2}\big)$ is positive for all $\alpha > -1$ and is also bounded. $\square$

Indeed, as now shown, asymptotically, $\mathcal{D}_n^i$ in (4.24) has full probability.

Lemma 4.4 *Let $2 \le i \le m-1$, then $\lim_{n\to+\infty} \mathbb{P}\big((\mathcal{D}_n^i)^c\big) = 0$.*

Proof. Clearly,

$$\begin{aligned}\mathbb{P}\big((\mathcal{D}_n^i)^c\big) &\le \sum_{(k_1,\dots,k_{i-1})\in\mathcal{C}_n^{i-1}} \mathbb{P}\big((D_n^i(k_1,\dots,k_{i-1}))^c\big)\\ &\le n^{i-1} \max_{(k_1,\dots,k_{i-1})\in\mathcal{C}_n^{i-1}} \mathbb{P}\big((D_n^i(k_1,\dots,k_{i-1}))^c\big).\end{aligned}$$

Therefore, to prove the lemma, it is enough to show that:

$$\lim_{n\to+\infty} n^{i-1} \max_{(k_1,\dots,k_{i-1})\in\mathcal{C}_n^{i-1}} \mathbb{P}\big((D_n^i(k_1,\dots,k_{i-1}))^c\big) = 0. \tag{4.27}$$

Now, for each $2 \le i \le m-1$, Propositions 3.1 and 3.2 assert that,

$$N_i^* = N_i^*(k_1,\dots,k_{i-1}) = \sum_{j=1}^{i-1} N_{i,j}^*,$$

where the $(N_{i,j}^*)_{1\le j\le i-1}$ are independent and with probability generating function

$$\mathbb{E}\left[x^{N_{i,j}^*}\right] = \left(\frac{1}{2-x}\right)^{k_j}.$$

Next,

$$\begin{aligned}\mathbb{P}\left(D_n^i(k_1,\dots,k_{i-1}))^c\right) &= \mathbb{P}\big(|N_i^* - \mathbb{E}[N_i^*]| > x_n\big)\\ &= \mathbb{P}\left(\sum_{j=1}^{i-1}\big(N_{i,j}^* - k_j\big) > x_n\right) + \mathbb{P}\left(\sum_{j=1}^{i-1}\big(k_j - N_{i,j}^*\big) > x_n\right). \quad (4.28)\end{aligned}$$

The first term in (4.28) is bounded by $\Theta^r_{k_1+\dots+k_{i-1}}(x_n)$, where

$$\begin{aligned}\Theta_k^r(x) &:= \min_{t>0}\Big(\exp\big(-\big(t(x+k) + k\ln(2-e^t)\big)\big)\Big) \quad (4.29)\\ &= \frac{(x+2k)^{x+2k}}{(2x+2k)^{x+k}(2k)^k}, \quad (4.30)\end{aligned}$$

since the minimization in (4.29) occurs at $t = \ln\big((2x+2k)/(x+2k)\big)$.

The second term in (4.28) is bounded by $\Theta^l_{k_1+\dots+k_{i-1}}(x_n)$, where

$$\begin{aligned}\Theta_k^l(x) &:= \min_{t>0}\Big(\exp\big(-\big(t(x-k) + k\ln(2-e^{-t})\big)\big)\Big) \quad (4.31)\\ &= \frac{(2k-x)^{2k-x}}{(2k-2x)^{k-x}(2k)^k}, \quad (4.32)\end{aligned}$$

observing that, for $x \le k$, the minimization in (4.31) occurs at $t = \ln\big((2k-x)/(2k-2x)\big)$. From the previous bounds and (4.28), it is clear that (4.27) will follow from

$$\lim_{n\to+\infty} n^{i-1} \max_{(k_1,\dots,k_{i-1})\in\mathcal{C}_n^{i-1}} \Theta^{\bullet}_{k_1+\cdots+k_{i-1}}(x_n) = 0, \tag{4.33}$$

for $\bullet \in \{l, r\}$. To obtain such as limit, we make use of Lemma 4.3, with $x = x_n = n^\beta$, $\beta \in (1/2, 1)$, noting also that $\mathcal{C}_n^{i-1} \subset \big\{(k_1,\dots,k_{i-1}) : k_1+\cdots+k_{i-1} \le \sum_{j=1}^{i-1}\mathbb{E}[N_j] + (i-1)\sqrt{n}\ln n\big\} \subset \big\{(k_1,\dots,k_{i-1}) : k_1+\cdots+k_{i-1} \le (i-1)(\max_{j=1,\dots,i-1} p_j n + \sqrt{n}\ln n)\big\}$.

First, for $\bullet = r$, when $k_1+\cdots+k_{i-1} \le x_n$, (4.25) writes as

$$\begin{aligned}\Theta^r_{k_1+\cdots+k_{i-1}}(x_n) &\le C\exp\Big(-c(k_1+\cdots+k_{i-1})\min\Big(\frac{x_n}{k_1+\cdots+k_{i-1}}, \Big(\frac{x_n}{k_1+\cdots+k_{i-1}}\Big)^2\Big)\Big)\\ &= C\exp(-cx_n),\end{aligned}$$

so that

$$n^{i-1}\max_{k_1+\cdots+k_{i-1}\le x_n} \Theta^r_{k_1+\cdots+k_{i-1}}(x_n) \le Cn^{i-1}e^{-cn^\beta} \to 0, \quad n\to+\infty,$$

where above, and below, C is a finite positive constant whose value might change from a line to another. For $x_n \le k_1+\cdots+k_{i-1} \le (i-1)(n\max_{j=1,\dots,i-1} p_j + \sqrt{n}\ln n) = (i-1)(n/m+\sqrt{n}\ln n)$, (4.25) writes as

$$\begin{aligned}\Theta^r_{k_1+\cdots+k_{i-1}}(x_n) &\le C\exp\Big(-c(k_1+\cdots+k_{i-1})\min\Big(\frac{x_n}{k_1+\cdots+k_{i-1}}, \Big(\frac{x_n}{k_1+\cdots+k_{i-1}}\Big)^2\Big)\Big)\\ &= C\exp\Big(-c\frac{x_n^2}{k_1+\cdots+k_{i-1}}\Big)\\ &\le C\exp\left(-c\frac{x_n^2}{(i-1)(n/m+\sqrt{n}\ln n)}\right),\end{aligned}$$

so that

$$\begin{aligned}&n^{i-1}\max_{x_n\le k_1+\cdots+k_{i-1}\le(i-1)(n/m+\sqrt{n}\ln n)} \Theta^r_{k_1+\cdots+k_{i-1}}(x_n)\\ &\qquad\le n^{i-1}\exp\left(-c\frac{n^{2\beta}}{(i-1)(n/m+\sqrt{n}\ln n)}\right) \to 0, \quad n\to+\infty,\end{aligned}$$

guaranteeing (4.33) with $\bullet = r$.

Next, let $\bullet = l$ and consider the following three cases: $k_1+\cdots+k_{i-1} \le x_n/2$, $x_n/2 \le k_1+\cdots+k_{i-1} \le x_n$ and $x_n \le k_1+\cdots+k_{i-1} \le (i-1)(n\max_{j=1,\dots,i-1} p_j + \sqrt{n}\ln n) = (i-1)(n/m+\sqrt{n}\ln n)$. When $k_1+\cdots+k_{i-1} \le x_n/2$, (4.31) ensures that for all $t>0$:

$$\begin{aligned}\Theta^l_{k_1+\cdots+k_{i-1}}(x_n) &\le \exp\Big(t(k_1+\cdots+k_{i-1}-x_n) - (k_1+\cdots+k_{i-1})\ln(2-e^{-t})\Big)\\ &\le \exp\Big(-\frac{t}{2}x_n\Big).\end{aligned}\tag{4.34}$$

When $x_n/2 \le k_1+\cdots+k_{i-1} \le x_n$, (4.31) ensures that for all $t>0$:

$$\begin{aligned}\Theta^l_{k_1+\cdots+k_{i-1}}(x_n) &\le \exp\Big(t(k_1+\cdots+k_{i-1}-x_n)-(k_1+\cdots+k_{i-1})\ln(2-e^{-t})\Big)\Big) \\ &\le \exp\Big(-\frac{x_n}{2}\ln(2-e^{-t})\Big). \qquad (4.35)\end{aligned}$$

When $x_n \le k_1+\cdots+k_{i-1} \le (i-1)(n/m+\sqrt{n}\ln n)$, (4.32) and (4.25) in Lemma 4.3 ensure that:

$$\begin{aligned}\Theta^l_{k_1+\cdots+k_{i-1}}(x_n) &\le C\exp\Big(-c(k_1+\cdots+k_{i-1})\min\Big(\frac{x_n}{k_1+\cdots+k_{i-1}},\Big(\frac{x_n}{k_1+\cdots+k_{i-1}}\Big)^2\Big)\Big) \\ &= C\exp\Big(-c\frac{x_n^2}{k_1+\cdots+k_{i-1}}\Big) \\ &\le C\exp\Big(-c\frac{n^{2\beta}}{(i-1)(n/m+\sqrt{n}\ln n)}\Big). \qquad (4.36)\end{aligned}$$

Gathering together the bounds (4.34), (4.35) and (4.36) proves (4.33), for $\bullet = l$. Combining this last fact with the corresponding result for $\bullet = r$, and via (4.33) and (4.27), proves Lemma 4.4. $\square$

Now, thanks to Lemma 4.4, to prove the convergence to zero, as $n\to+\infty$, of the first term on the right-hand side of (4.21), it is enough to prove the same result for

$$\mathbb{P}\left(\left\{\max_{\boldsymbol{k}\in\mathcal{C}^{**}_{n,i}}\left|\sum_{j=\ell_i+1}^{\ell_i+N_i^*-(k_1+\cdots+k_{i-1})} Z_j^{(i)}\right| \ge \varepsilon\sqrt{n}\right\}\cap\mathcal{D}_n^i\right), \qquad (4.37)$$

where the $Z_j^{(i)}$ are given in (4.6), i.e., $Z_j^{(i)} = \big(N_m^{T_i^{j-1},T_i^j}-1\big)/\sqrt{2}$, $i=1,\dots,m-1$, $j\ge 1$. Our next elementary lemma, the ultimate before closing this section, provides tail estimates on the partial sums of the Z_j (omitting the indices i for a while).

Lemma 4.5 *Let $(Z_j)_{j\ge1}$ be iid random variables as in (4.6). Then, for suitable positive and finite constants c and C, all $x>0$, and positive integer n,*

$$\begin{aligned}\mathbb{P}\Big(\sum_{j=1}^n Z_j \ge x\Big) &\le \min_{t>0}\Big(\exp\Big(-\Big(t(x\sqrt{2}+n)+n\ln(2-e^t)\Big)\Big)\Big) = \Theta^r_n(x\sqrt{2}), \quad (4.38)\\ \mathbb{P}\Big(\sum_{j=1}^n Z_j \ge x\Big) &\le C\exp\Big(-c\min\Big(\frac{x}{n},\Big(\frac{x}{n}\Big)^2\Big)\Big), \quad (4.39)\\ \mathbb{P}\Big(\sum_{j=1}^n Z_j \le -x\Big) &\le \min_{t>0}\Big(\exp\Big(-\Big(t(x\sqrt{2}-n)+n\ln(2-e^{-t})\Big)\Big)\Big) = \Theta^l_n(x\sqrt{2}), \quad (4.40)\\ \mathbb{P}\Big(\sum_{j=1}^n Z_j \le -x\Big) &\le C\exp\Big(-c\min\Big(\frac{x}{n},\Big(\frac{x}{n}\Big)^2\Big)\Big), \quad \textit{for } x\le n. \quad (4.41)\end{aligned}$$

Proof. Recall from (4.6) that $Z_j = \big(N_m^{T_i^{j-1},T_i^j} - 1\big)/\sqrt{2}$, $i \neq m$, and from (3.2),

$$\mathbb{E}\Big[x^{N_m^{T_i^{j-1},T_i^j}}\Big] = \frac{1}{2-x}. \tag{4.42}$$

Hence, using the notation in (4.29),

$$\mathbb{P}\Big(\sum_{j=1}^{n} Z_j \geq x\Big) \leq \min_{t>0}\Big(e^{-t(x\sqrt{2}+n)}\ \mathbb{E}\Big[\exp\big(tN_m^{T_i^{j-1},T_i^j}\big)\Big]^n\Big) = \Theta_n^r(x\sqrt{2}),$$

and (4.39) follows from (4.29) and (4.30) in (the proof of) Lemma 4.4 (with its notation) and from (4.25) in Lemma 4.3. Similarly, using the notation in (4.31)

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j=1}^{n} Z_j \leq -x\Big) &= \mathbb{P}\left(\sum_{j=1}^{n}\Big(1 - N_m^{T_i^{j-1},T_i^j}\Big) \geq x\sqrt{2}\right)\\
&\leq \min_{t>0}\Big(e^{-t(x\sqrt{2}-n)}\ \mathbb{E}\Big[\exp\big(-tN_m^{T_i^{j-1},T_i^j}\big)\Big]^n\Big) = \Theta_n^l(x\sqrt{2}).
\end{aligned}$$

which is (4.40). As previously observed via (4.32), when $x \leq n$, the minimization for $\Theta_n^l(x)$ occurs at $t = \ln\big((2n-x)/(2n-2x)\big)$, and, once again, (4.25) on Lemma 4.3 ensure (4.41). □

We are now ready to move towards completing this section. From its very definition in (4.22),

$$\begin{aligned}
\mathcal{C}_{n,i}^{**} &= \mathcal{C}_n^{i-1} \cap \big\{\mathbb{E}[N_i^*] \leq \ell_i = k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - \mathbb{E}[N_i^*])\big\}\\
&\subset \left\{k_1 + \cdots + k_{i-1} \leq \sum_{j=1}^{i-1}\mathbb{E}[N_j] + (i-1)\sqrt{n}\ln n\right\}\\
&\qquad \cap \big\{\mathbb{E}[N_i^*] \leq \ell_i = k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - \mathbb{E}[N_i^*])\big\}\\
&\subset \left\{k_1 + \cdots + k_{i-1} \leq (i-1)(n\max_{j=1,\dots,i-1} p_j + \sqrt{n}\ln n)\right\}\\
&\qquad \cap \big\{\mathbb{E}[N_i^*] \leq \ell_i = k_1 + \cdots + k_i \leq \mathbb{E}[N_i] + \sqrt{n}\ln n - (N_i^* - \mathbb{E}[N_i^*])\big\}.
\end{aligned}$$

Therefore, recalling also from (3.26) that $\mathbb{E}[N_i^*] = k_1 + \cdots + k_{i-1}$, (4.37) is upper bounded by:

$$\begin{aligned}
&\mathbb{P}\left(\left\{\max_{\substack{k_1+\cdots+k_{i-1}\leq(i-1)(n/m+\sqrt{n}\ln n)\\ k_1+\cdots+k_{i-1}\leq \ell_i \leq \mathbb{E}[N_i]+\sqrt{n}\ln n-(N_i^*-(k_1+\cdots+k_{i-1}))}} \left|\sum_{j=\ell_i+1}^{\ell_i+N_i^*-(k_1+\cdots+k_{i-1})} Z_j\right| \geq \varepsilon\sqrt{n}\right\} \cap \mathcal{D}_n^i\right)\\
&\qquad \leq \mathbb{P}\left(\max_{\substack{k_1+\cdots+k_{i-1}\leq(i-1)(n/m+\sqrt{n}\ln n)\\ k_1+\cdots+k_{i-1}\leq \ell_i \leq \mathbb{E}[N_i]+\sqrt{n}\ln n+x_n}} \max_{|n_i|\leq x_n}\left|\sum_{j=\ell_i+1}^{\ell_i+n_i} Z_j\right| \geq \varepsilon\sqrt{n}\right) \quad \text{(recall (4.24))}
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\max_{\ell_i \leq \mathbb{E}[N_i]+\sqrt{n}\ln n + x_n} \max_{|n_i|\leq x_n} \left|\sum_{j=\ell_i+1}^{\ell_i+n_i} Z_j\right| \geq \varepsilon\sqrt{n}\right)\\
&\leq 3nx_n \max_{\substack{\ell_i \leq \mathbb{E}[N_i]+\sqrt{n}\ln n + x_n\\ |n_i|\leq x_n}} \mathbb{P}\left(\left|\sum_{j=\ell_i+1}^{\ell_i+n_i} Z_j\right| \geq \varepsilon\sqrt{n}\right)\\
&\leq 3nx_n \max_{\substack{\ell_i \leq \mathbb{E}[N_i]+\sqrt{n}\ln n + x_n\\ 0\leq n_i\leq x_n}} \left(\Theta^l_{n_i}(\varepsilon\sqrt{2n}) + \Theta^r_{n_i}(\varepsilon\sqrt{2n})\right), \qquad (4.43)
\end{aligned}
$$

where, in the next to last inequality, we used the usual (sharp in the iid case) bounding of the maximum via the number of terms times the maximal probability; while in the last one, $|n_i| \leq x_n$ was changed into $0 \leq n_i \leq x_n = n^\beta$, $1/2 < \beta < 1$.

Our final task is to show that

$$
\lim_{n\to+\infty} nx_n \max_{0\leq n_i\leq x_n} \Theta^\bullet_{n_i}(\varepsilon\sqrt{2n}) = 0, \qquad (4.44)
$$

for $\bullet \in \{l, r\}$. This relies again on Lemma 4.3 and Lemma 4.5. For $\bullet = r$, when $k < \varepsilon\sqrt{2n}$, (4.38) and (4.39) entail that,

$$
\Theta^r_k(\varepsilon\sqrt{2n}) \leq C\exp(-c\varepsilon\sqrt{2n}); \qquad (4.45)
$$

while, for $\varepsilon\sqrt{2n} \leq k \leq x_n$, they entail that,

$$
\Theta^r_k(\varepsilon\sqrt{2n}) \leq C\exp\big(-2c\varepsilon^2 n/k\big) \leq C\exp\big(-2c\varepsilon^2 n/x_n\big) = C\exp\big(-2c\varepsilon^2 n^{1-\beta}\big). \qquad (4.46)
$$

Therefore, for $\bullet = r$, (4.44) follows from (4.45) and (4.46). Let us now turn our attention to $\bullet = l$. When $\varepsilon\sqrt{2n} \leq k \leq x_n$, (4.41) entails that,

$$
\Theta^l_k(\varepsilon\sqrt{2n}) \leq C\exp\big(-2c\varepsilon^2 n/k\big) \leq C\exp\big(-2c\varepsilon^2 n/x_n\big) = C\exp\big(-2c\varepsilon^2 n^{1-\beta}\big). \qquad (4.47)
$$

For $k \leq \varepsilon\sqrt{n/2}$, (4.40) entails that, for any $t > 0$,

$$
\Theta^l_k(\varepsilon\sqrt{2n}) \leq \exp\Big(t(k-\varepsilon\sqrt{2n}) - k\ln(2-e^{-t})\Big)\Big) \leq \exp\Big(-\varepsilon t\sqrt{n/2}\Big). \qquad (4.48)
$$

For $\varepsilon\sqrt{n/2} \leq k \leq \varepsilon\sqrt{2n}$, (4.40) entails that, for any $t > 0$,

$$
\Theta^l_k(\varepsilon\sqrt{2n}) \leq \exp\Big(t(k-\varepsilon\sqrt{n/2}) - k\ln(2-e^{-t})\Big)\Big) \leq \exp\Big(-\varepsilon t\sqrt{n/2}\ln(2-e^{-t})\Big). \qquad (4.49)
$$

Therefore, for $\bullet = l$, (4.44) follows from (4.47), (4.48), and (4.49). Gathering all the intermediate results, for any $i = 2, \ldots, m-1$,

$$
\lim_{n\to+\infty} \mathbb{P}\left(\left\{\max_{\boldsymbol{k}\in\mathcal{C}^{**}_{n,i}} \left|\sum_{j=\ell_i+1}^{\ell_i+N^*_i-(k_1+\cdots+k_{i-1})} Z^{(i)}_j\right| \geq \varepsilon\sqrt{n}\right\} \cap \mathcal{D}^i_n\right) = 0,
$$

and therefore,

$$
\lim_{n\to+\infty} \mathbb{P}\left(\max_{\boldsymbol{k}\in\mathcal{C}^*} \left|\sum_{j=k_1+\cdots+k_i+1}^{N^*_i+k_i} \frac{Z^{(i)}_j}{\sqrt{n}}\right| \geq \varepsilon\right) = 0.
$$

The goal of this section has thus been achieved: the quantities (4.13) and (4.14) have the same weak limit.

4.3 The Constraints

To deal with the third heuristic limit, we now need to obtain the convergence of the random set of constraints towards a deterministic set of constraints. This fact will follow from the various reductions obtained to date as well as new arguments developed from now on. To start with, let us recall two elementary facts about convergence in distribution.

The first fact asserts that if $(f_n)_{1\le n\le\infty}$ is a sequence of Borel functions such that $x_n \to x_\infty$ implies that $f_n(x_n) \to f_\infty(x_\infty)$, and if $(X_n)_{n\ge1}$ is a sequence of random variables such that $X_n \Rightarrow X_\infty$, then $f_n(X_n) \Rightarrow f_\infty(X_\infty)$. Indeed, via the Skorohod representation theorem, there exists a probability space and random variables Y_n, $1 \le n \le \infty$, such that $Y_n \stackrel{\mathcal{L}}{=} X_n$, $1 \le n \le \infty$, and $Y_n \to Y_\infty$ with probability one. But, by hypothesis, $f_n(Y_n) \to f_\infty(Y_\infty)$, with probability one. Therefore $f_n(X_n) \Rightarrow f_\infty(X_\infty)$.

The second elementary fact is as follows: Let $(X_n)_{n\ge1}$ be a sequence of random variables such that $X_n^{\pm} \Rightarrow Y$, then $X_n \Rightarrow Y$, where $x^+ = \max(x,0)$ and $x^- = \min(x,0)$. Indeed, $\mathbb{P}(X_n^+ \le x) \le \mathbb{P}(X_n \le x) \le \mathbb{P}(X_n^- \le x)$, for all $x \in \mathbb{R}$.

Using these two elementary facts, let us return to our derandomization problem. Recalling (4.14), and using the polygonal structure of the B_n^X and B_n^Y, we have

$$M_n \;:=\; \max_{\boldsymbol{k}\in\widehat{\mathcal{C}}_n}\left(F_X\left(B_n^X,\frac{\boldsymbol{k}}{n}\right)\wedge F_Y\left(B_n^Y,\frac{\boldsymbol{k}}{n}\right)\right),$$

where

$$\widehat{\mathcal{C}}_n = \Big\{\boldsymbol{k} = (k_i)_{1\le i\le m-1} : \forall i = 1,\dots,m-1,\ 0 \le k_i \le n$$
$$0 \le k_i \le \big(N_i(X) - N_i^*(X)\big) \wedge \big(N_i(Y) - N_i^*(Y)\big)\Big\}, \tag{4.50}$$

and where

$$F_X\big(\mathbf{x},\mathbf{t}\big) \;=\; \frac{1}{m}\sum_{i=1}^{m-1} x_i(p_i(X)) - \sum_{i=1}^{m-1}\left(x_i\Big(\sum_{j=1}^{i} t_j\Big) - x_i\Big(\sum_{j=1}^{i-1} t_j\Big)\right), \tag{4.51}$$

$$F_Y\big(\mathbf{x},\mathbf{t}\big) \;=\; \frac{1}{m}\sum_{i=1}^{m-1} x_i(p_i(Y)) - \sum_{i=1}^{m-1}\left(x_i\Big(\sum_{j=1}^{i} t_j\Big) - x_i\Big(\sum_{j=1}^{i-1} t_j\Big)\right), \tag{4.52}$$

for $\mathbf{x} = (x_1,\dots,x_{m-1}) \in \big(C_0([0,1])\big)^{m-1}$ and $\mathbf{t} = (t_1,\dots,t_{m-1}) \in [0,1]^{m-1}$. Now, let

$$\widehat{\mathcal{C}}_n^{\pm} = \left\{\boldsymbol{k} = (k_i)_{1\le i\le m-1} : \forall\ i = 1,\dots,m-1, 0 \le k_i \le n \text{ and } \sum_{j=1}^{i}\frac{k_j}{n} \le p_i \pm 2x_n\right\}, \tag{4.53}$$

with $x_n = n^\beta$, $\beta \in (1/2,1)$ as in (4.23), and let

$$M_n^{\pm} = \max_{\boldsymbol{k}\in\widehat{\mathcal{C}}_n^{\pm}}\left(F_X\left(B_n^X,\frac{\boldsymbol{k}}{n}\right)\wedge F_Y\left(B_n^Y,\frac{\boldsymbol{k}}{n}\right)\right). \tag{4.54}$$

Since

$$N_i(X) - N_i^*(X) \;=\; np_i - \sum_{j=1}^{i-1} k_i + \Big(\big(N_i(X) - \mathbb{E}[N_i(X)]\big) - \big(N_i^*(X) - \mathbb{E}[N_i^*(X)]\big)\Big),$$

with a similar statement replacing X by Y, the condition

$$k_i \le \big(N_i(X) - N_i^*(X)\big) \wedge \big(N_i(Y) - N_i^*(Y)\big),$$

in (4.50), writes as $\sum_{j=1}^{i} k_i/n \le p_i + R_n^i(X,Y)/n$ where

$$\begin{aligned} R_n^i(X,Y) \;=\; & \Big(\big(N_i(X) - \mathbb{E}[N_i(X)]\big) - \big(N_i^*(X) - \mathbb{E}[N_i^*(X)]\big)\Big) \\ & \wedge\Big(\big(N_i(Y) - \mathbb{E}[N_i(Y)]\big) - \big(N_i^*(Y) - \mathbb{E}[N_i^*(Y)]\big)\Big). \end{aligned} \tag{4.55}$$

Now let

$$\mathcal{K}_n = \bigcap_{i=1}^{m-1} \big\{|N_i - \mathbb{E}[N_i]| < x_n\big\} \cap \mathcal{D}_n^i$$

with $\mathcal{D}_n^i$ defined in (4.24). From (4.11) and Lemma 4.4, we have $\lim_{n\to+\infty} \mathbb{P}\big(\mathcal{K}_n^c\big) = 0$ and, on $\mathcal{K}_n$, $R_n^i(X,Y) \le 2x_n$, for all $1 \le i \le m$. Therefore, $\widehat{\mathcal{C}}_n^- \subset \widehat{\mathcal{C}}_n \subset \widehat{\mathcal{C}}_n^+$ on $\mathcal{K}_n$, and

$$M_n^- \le M_n \le M_n^+. \tag{4.56}$$

Clearly,

$$M_n^{\pm} = \max_{\mathbf{t}\in\mathcal{C}_n^{\pm}} \left(F_X\left(B_n^X, \mathbf{t}\right) \wedge F_Y\left(B_n^Y, \mathbf{t}\right)\right), \tag{4.57}$$

where now

$$\mathcal{C}_n^{\pm} = \left\{\mathbf{t} = (t_i)_{1\le i\le m-1} \in [0,1]^{m-1} : \forall\ i = 1,\dots,m-1, \sum_{j=1}^{i} t_j \le p_i \pm 2\frac{x_n}{n}\right\}. \tag{4.58}$$

Next,

$$\begin{aligned} \mathbb{P}\big(M_n \le x\big) &\le \mathbb{P}\Big(\{M_n \le x\} \cap \mathcal{K}_n\Big) + \mathbb{P}\big(\mathcal{K}_n^c\big) \\ &\le \mathbb{P}\big(\{M_n^- \le x\} \cap \mathcal{K}_n\big) + \mathbb{P}\big(\mathcal{K}_n^c\big) \\ &\le \mathbb{P}\big(M_n^- \le x\big) + \mathbb{P}\big(\mathcal{K}_n^c\big), \end{aligned}$$

therefore

$$\limsup_{n\to+\infty} \mathbb{P}(M_n \le x) \le \limsup_{n\to+\infty} \mathbb{P}(M_n^- \le x). \tag{4.59}$$

Similarly,

$$\mathbb{P}\big(M_n \le x\big) = \mathbb{P}\Big(\{M_n \le x\} \cap \mathcal{K}_n\Big) + \mathbb{P}\Big(\{M_n \le x\} \cap \mathcal{K}_n^c\Big)$$

$$\geq \mathbb{P}\Big(\{M_n^+ \leq x\} \cap \mathcal{K}_n\Big)$$
$$\geq \mathbb{P}\big(M_n^+ \leq x\big) - \mathbb{P}\big(\mathcal{K}_n^c\big),$$

and therefore

$$\liminf_{n\to+\infty} \mathbb{P}(M_n \leq x) \geq \liminf_{n\to+\infty} \mathbb{P}\big(M_n^+ \leq x\big). \tag{4.60}$$

Combining (4.59) and (4.60) with the second elementary fact described above, our goal is now to show that convergence in distribution of both M_n^+ and M_n^- towards

$$M_\infty = \max_{\mathbf{t}\in\mathcal{V}} \big(F_X(B^X,\mathbf{t}) \wedge F_Y(B^Y,\mathbf{t})\big), \tag{4.61}$$

holds true, where

$$\mathcal{V} := \mathcal{V}(p_1,\ldots,p_{m-1}) = \left\{\mathbf{t} = (t_j)_{1\leq j\leq m-1} \in [0,1]^{m-1} : \forall i = 1,\ldots,m-1, \sum_{j=1}^{i} t_j \leq p_i\right\}.$$

To do so, first note that by Donsker's theorem $(B_n^X, B_n^Y) \Rightarrow (B^X, B^Y)$ and we now wish to apply the first elementary fact, recalled above, to the functions

$$f_n^\pm(\mathbf{x},\mathbf{y}) = \max_{\mathbf{t}\in\mathcal{C}_n^\pm} \big(F_X\left(\mathbf{x},\mathbf{t}\right) \wedge F_Y\left(\mathbf{y},\mathbf{t}\right)\big), \tag{4.62}$$

and

$$f_\infty(\mathbf{x},\mathbf{y}) = \max_{\mathbf{t}\in\mathcal{V}} \big(F_X(\mathbf{x},\mathbf{t}) \wedge F_Y(\mathbf{y},\mathbf{t})\big). \tag{4.63}$$

With these notations, $M_n^\pm = f_n^\pm(B_n^X, B_n^Y)$ and $M_\infty = f_\infty(B^X, B^Y)$. In other words, we wish to show that $(\mathbf{x}_n,\mathbf{y}_n) \to (\mathbf{x},\mathbf{y})$ in $(C_0([0,1]))^{m-1}$ implies that $f_n(\mathbf{x}_n,\mathbf{y}_n) \to f_\infty(\mathbf{x},\mathbf{y})$. To start with,

$$|f_n^\pm(\mathbf{x}_n,\mathbf{y}_n) - f_\infty(\mathbf{x},\mathbf{y})| \leq |f_n^\pm(\mathbf{x}_n,\mathbf{y}_n) - f_n^\pm(\mathbf{x},\mathbf{y})| + |f_n^\pm(\mathbf{x},\mathbf{y}) - f_\infty(\mathbf{x},\mathbf{y})|, \tag{4.64}$$

and we continue by estimating $|f_n^\pm(\mathbf{x}_n,\mathbf{y}_n) - f_n^\pm(\mathbf{x},\mathbf{y})|$. But,

$$\begin{aligned}
&|f_n^\pm(\mathbf{x}_n,\mathbf{y}_n) - f_n^\pm(\mathbf{x},\mathbf{y})| \\
&\quad\leq \max_{\mathbf{t}\in\mathcal{C}_n^\pm}\Big|\Big(F_X\left(\mathbf{x}_n,\mathbf{t}\right) \wedge F_Y\left(\mathbf{y}_n,\mathbf{t}\right)\Big) - \Big(F_X\left(\mathbf{x},\mathbf{t}\right) \wedge F_Y\left(\mathbf{y},\mathbf{t}\right)\Big)\Big| \\
&\quad\leq \max_{\mathbf{t}\in\mathcal{C}_n^\pm} \max\Big(\left|F_X\left(\mathbf{x}_n,\mathbf{t}\right) - F_X\left(\mathbf{x},\mathbf{t}\right)\right|, \left|F_Y\left(\mathbf{y}_n,\mathbf{t}\right) - F_Y\left(\mathbf{y},\mathbf{t}\right)\right|\Big) \qquad (4.65) \\
&\quad\leq C \max_{\mathbf{t}\in\mathcal{C}_n^\pm} \max\Big(\left|\mathbf{x}_n(\mathbf{t}) - \mathbf{x}(\mathbf{t})\right|, \left|\mathbf{y}_n(\mathbf{t}) - \mathbf{y}(\mathbf{t})\right|\Big), \qquad (4.66)
\end{aligned}$$

making use of Lemma 4.1 in (4.65) and by the linearity of both F_X and F_Y with respect to their first argument in (4.66) and where, further, C is a finite positive constant. Therefore,

$$|f_n^\pm(\mathbf{x}_n,\mathbf{y}_n) - f_n^\pm(\mathbf{x},\mathbf{y})| \leq C \max\big(\|\mathbf{x}_n - \mathbf{x}\|_\infty, \|\mathbf{y}_n - \mathbf{y}\|_\infty\big),$$

and so if $(\mathbf{x}_n, \mathbf{y}_n) \to (\mathbf{x}, \mathbf{y})$, it follows that $f_n^{\pm}(\mathbf{x}_n, \mathbf{y}_n) - f_n^{\pm}(\mathbf{x}, \mathbf{y}) \to 0$.

In order to complete the proof of $M_n^{\pm} \Rightarrow M_\infty$ and thus that of $M_n \Rightarrow M_\infty$, let us now estimate the right-most expression in (4.64).

At first, note that $\mathcal{C}_n^- \subset \mathcal{V} \subset \mathcal{C}_n^+$, hence

$$f_n^-(\mathbf{x}, \mathbf{y}) \le f_\infty(\mathbf{x}, \mathbf{y}) \le f_n^+(\mathbf{x}, \mathbf{y}). \tag{4.67}$$

Next, from (4.62) and (4.63), set $f_n^+(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{t}\in\mathcal{C}_n^+} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$, and $f_\infty(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{t}\in\mathcal{V}} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$, where $\theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) = F_X(\mathbf{x}, \mathbf{t}) \wedge F_Y(\mathbf{y}, \mathbf{t})$. Since $\mathcal{C}_n^- \subset \mathcal{C}_{n+1}^-$, for $n \ge 1$, it follows (as shown next) that $f_n^-(\mathbf{x}, \mathbf{y}) \to \max_{\mathbf{t}\in\bigcup_{n\ge1}\mathcal{C}_n^-} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$. Indeed, $\lim_{n\to+\infty} f_n^-(\mathbf{x}, \mathbf{y}) \le \max_{\mathbf{t}\in\bigcup_{n\ge1}\mathcal{C}_n^-} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$ and if the previous inequality were strict, there would be $K \in (0, +\infty)$ such that

$$\max_{\mathbf{t}\in\mathcal{C}_n^-} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) \le K < \max_{\mathbf{t}\in\bigcup_{n\ge1}\mathcal{C}_n^-} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}).$$

The left-hand side inequality implies that for all $n \ge 1$, and $\mathbf{t} \in \mathcal{C}_n^-$, $\theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) \le K$, contradicting the right-hand side inequality.

Since $\mathcal{C}_n^+ \supset \mathcal{C}_{n+1}^+$, for $n \ge 1$, it also follows that $f_n^+(\mathbf{x}, \mathbf{y}) \to \max_{\mathbf{t}\in\bigcap_{n\ge1}\mathcal{C}_n^+} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$. Indeed, we have $\lim_{n\to+\infty} f_n^+(\mathbf{x}, \mathbf{y}) \ge \max_{\mathbf{t}\in\bigcap_{n\ge1}\mathcal{C}_n^-} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t})$ and if the previous inequality were strict, there would be $K \in (0, +\infty)$ such that

$$\max_{\mathbf{t}\in\mathcal{C}_n^+} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) \ge K > \max_{\mathbf{t}\in\bigcap_{n\ge1}\mathcal{C}_n^+} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}).$$

The left-hand side inequality implies that for any $n \ge 1$, there exists $\mathbf{t}_n \in \mathcal{C}_n^+$ with $\theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}_n) \ge K$. Up to a subsequence $\mathbf{t}_n \to \mathbf{t}^* \in \bigcap_{n\ge1}\mathcal{C}_n^+$ and by continuity of $\theta_{\mathbf{x},\mathbf{y}}$, $\theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}^*) \ge K$, which is inconsistent with the previous right-hand side inequality.

Finally, since $\bigcup_{n\ge1}\mathcal{C}_n^- = \mathcal{V}^\circ$, is the interior of $\mathcal{V}$, and since $\bigcap_{n\ge1}\mathcal{C}_n^+ = \overline{\mathcal{V}} = \mathcal{V}$, is the closure of $\mathcal{V}$, we have

$$\lim_{n\to+\infty} f_n^-(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{t}\in\mathcal{V}^\circ} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) \le f_\infty(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{t}\in\mathcal{V}} \theta_{\mathbf{x},\mathbf{y}}(\mathbf{t}) = \lim_{n\to+\infty} f_n^+(\mathbf{x}, \mathbf{y}). \tag{4.68}$$

It remains to show that the maximum of $\theta_{\mathbf{x},\mathbf{y}}$ on $\mathcal{V}$ is attained on $\mathcal{V}^\circ$ for $\mathbb{P}_{(B^X,B^Y)}$-almost all $(\mathbf{x}, \mathbf{y})$, i.e., that

$$\mathbb{P}\left(\max_{\mathbf{t}\in\mathcal{V}(1/m,\dots,1/m)^\circ} \theta_{B^X,B^Y}(\mathbf{t}) = \max_{\mathbf{t}\in\mathcal{V}(1/m,\dots,1/m)} \theta_{B^X,B^Y}(\mathbf{t})\right) = 1. \tag{4.69}$$

With (4.69), (4.68) entails $\lim_{n\to+\infty} f_n^{\pm}(\mathbf{x}, \mathbf{y}) = f_\infty(\mathbf{x}, \mathbf{y})$ for $\mathbb{P}_{(B^X,B^Y)}$-almost all $(\mathbf{x}, \mathbf{y})$, i.e., the right-most expression in (4.64) converges to 0 and, as previously explained, this gives $M_n^{\pm} \Rightarrow M_\infty$ and $M_n \Rightarrow M_\infty$.

In order to complete (4.69), we anticipate, in the second equality below, on Section 4.4 in which parameters are changes via: $s_1 = u_1, s_1 + s_2 = u_2, \dots, s_1 + \cdots + s_{m-1} = u_{m-1}$ and where we prove that

$$\left(\theta_{B^X,B^Y}(\mathbf{t})\right)_{\mathbf{t}\in\mathcal{V}(1/m,\dots,1/m)} \overset{\mathcal{L}}{=} \frac{1}{\sqrt{m}}\left(\theta_{B^X,B^Y}(\mathbf{s})\right)_{\mathbf{s}\in\mathcal{V}(1,\dots,1)} = \frac{1}{\sqrt{2m}}\left(\widetilde{\theta}_{B_1,B_2}(\mathbf{u})\right)_{\mathbf{u}\in\mathcal{W}_m(1)},$$

where $\mathcal{W}_m(1) = \{0 = u_0 \le u_1 \le \cdots \le u_{m-1} \le u_m = 1\}$,

$$\begin{aligned}\widetilde{\theta}_{B_1,B_2}(\mathbf{u}) \;&=\; \left(-\frac{1}{m}\;\sum_{i=1}^{m} B_1^{(i)}(1) + \sum_{i=1}^{m}\left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right)\right) \\ &\qquad \wedge \left(-\frac{1}{m}\sum_{i=1}^{m} B_2^{(i)}(1) + \sum_{i=1}^{m}\left(B_2^{(i)}(u_i) - B_2^{(i)}(u_{i-1})\right)\right),\end{aligned} \tag{4.70}$$

and where B_1, B_2 are two independent, standard, m-dimensional Brownian on $[0,1]$. The property (4.69) is thus equivalent to

$$\mathbb{P}\left(\max_{\mathbf{u}\in\mathcal{W}_m(1)^\circ} \widetilde{\theta}_{B_1,B_2}(\mathbf{u}) = \max_{\mathbf{u}\in\mathcal{W}_m(1)} \widetilde{\theta}_{B_1,B_2}(\mathbf{u})\right) = 1. \tag{4.71}$$

The advantage of (4.71) over (4.69) is that the former involves two standard Brownian motions each one having *independent* coordinates. Roughly speaking, the property (4.71) should be derived from the following observation: when $\mathbf{u} \in \partial\mathcal{W}_m(1)$, then $u_k = u_{k+1}$, for some index k, and for such a $\mathbf{u}$, the sum $\sum_{i=1}^m \left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right)$ contains only $m-1$ terms. Letting $\mathbf{u}_\varepsilon$ be given by

$$u_{\varepsilon,i} = u_i, \quad i \neq k+1, \qquad \text{and} \qquad u_{\varepsilon,k+1} = u_k + \varepsilon,$$

we have

$$\begin{aligned}&\sum_{i=1}^{m}\left(B_1^{(i)}(u_{\varepsilon,i}) - B_1^{(i)}(u_{\varepsilon,i-1})\right) \\ &\quad = \sum_{i=1}^{m}\left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right) + \left(B_1^{(k+1)}(u_k+\varepsilon) - B_1^{(k+1)}(u_k)\right) \\ &\qquad\quad + \left(B_1^{(k+2)}(u_k) - B_1^{(k+2)}(u_k+\varepsilon)\right).\end{aligned}$$

The terms $\left(B_1^{(k+1)}(u_k+\varepsilon) - B_1^{(k+1)}(u_k)\right)$ and $\left(B_1^{(k+2)}(u_k) - B_1^{(k+2)}(u_k+\varepsilon)\right)$ are independent of $\sum_{i=1}^m \left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right)$ and from standard properties of Brownian motion, almost surely, the sum $\left(B_1^{(k+1)}(u_k+\varepsilon) - B_1^{(k+1)}(u_k)\right) + \left(B_1^{(k+2)}(u_k) - B_1^{(k+2)}(u_k+\varepsilon)\right)$ takes positive value for arbitrarily small $\varepsilon > 0$. Since the same is true for the second term in (4.70) relative to B_2, it follows that in the vicinity of each $\mathbf{u} \in \partial\mathcal{W}_m(1)$, there is $\mathbf{u}_\varepsilon \in \mathcal{W}_m(1)$ with $\widetilde{\theta}_{B_1,B_2}(\mathbf{u}_\varepsilon) > \widetilde{\theta}_{B_1,B_2}(\mathbf{u})$. Therefore, $\max_{\mathbf{u}\in\mathcal{W}_m(1)} \widetilde{\theta}_{B_1,B_2}(\mathbf{u})$ is attained in $\mathcal{W}_m(1)^\circ$, and so both (4.71) and (4.69) hold true, leading to $M_n \Rightarrow M_\infty$.

4.4 Final Step: A Linear Transformation

By combining the results of the last three subsections, we proved that

$$\frac{\mathrm{LCI}_n - n/m}{\sqrt{2n}} \Rightarrow \max_{\mathcal{V}(1/m,\ldots,1/m)} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}\left(\frac{1}{m}\right) - \sum_{i=1}^{m-1}\left(B^{i,X}\left(\sum_{j=1}^{i} t_j\right) - B^{i,X}\left(\sum_{j=1}^{i-1} t_j\right)\right)\right),$$

$$\frac{1}{m}\sum_{i=1}^{m-1} B^{i,Y}\left(\frac{1}{m}\right) - \sum_{i=1}^{m-1}\left(B^{i,Y}\left(\sum_{j=1}^{i} t_j\right) - B^{i,Y}\left(\sum_{j=1}^{i-1} t_j\right)\right)\Bigg) \quad (4.72)$$

where the maximum is taken over $\boldsymbol{t} = (t_1, \dots, t_{m-1}) \in \mathcal{V}(1/m, \dots, 1/m)$. Now, via the linear transformations of the parameters given by $u_i = m\sum_{j=1}^{i} t_j$, $i = 1, \dots, m-1$, $u_0 = t_0 = 0$, and Brownian scaling, the right-hand side of (4.72) becomes equal to:

$$\frac{1}{\sqrt{m}} \max_{0=u_0\le u_1\le\cdots\le u_{m-1}\le 1} \min\left(\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}(1) - \sum_{i=1}^{m-1}\left(B^{i,X}(u_i) - B^{i,X}(u_{i-1})\right), \right.$$
$$\left. \frac{1}{m}\sum_{i=1}^{m-1} B^{i,Y}(1) - \sum_{i=1}^{m-1}\left(B^{i,Y}(u_i) - B^{i,Y}(u_{i-1})\right)\right). \quad (4.73)$$

Next, for all $t \in [0,1]$ and $i = 1, \dots, m-1$, let us introduce the following two pointwise linear transformations:

$$B^{i,X}(t) = \frac{B_1^{(m)}(t) - B_1^{(i)}(t)}{\sqrt{2}},$$
$$B^{i,Y}(t) = \frac{B_2^{(m)}(t) - B_2^{(i)}(t)}{\sqrt{2}},$$

where B_1 and B_2 are two, standard, m-dimensional Brownian motion on $[0,1]$. Clearly $(B^{1,X}(t), \dots, B^{m-1,X}(t))_{0\le t\le 1}$ has the correct covariance matrix with diagonal entries given by (3.3) and off-diagonal ones given by (3.6) (in the uniform case they are respectively 1 on the diagonal and 1/2 elsewhere), and similarly, replacing X by Y. Moreover,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m-1} B^{i,X}(1) - \sum_{i=1}^{m-1}\left(B^{i,X}(u_i) - B^{i,X}(u_{i-1})\right) \\
&= -\frac{1}{\sqrt{2}m}\left(\sum_{i=1}^{m} B_1^{(i)}(1)\right) + \frac{1}{\sqrt{2}} B_1^{(m)}(1) \\
&\quad -\frac{1}{\sqrt{2}}\sum_{i=1}^{m-1}\left(B_1^{(m)}(u_i) - B_1^{(m)}(u_{i-1})\right) + \frac{1}{\sqrt{2}}\sum_{i=1}^{m-1}\left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right) \\
&= \frac{1}{\sqrt{2}}\left(-\frac{1}{m}\sum_{i=1}^{m} B_1^{(i)}(1) + (B_1^{(m)}(1) - B_1^{(m)}(u_{m-1})) + \sum_{i=1}^{m-1}\left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right)\right). \quad (4.74)
\end{aligned}$$

Finally, with the help of (4.74) (and the corresponding identity for Y), (4.73) becomes:

$$\frac{1}{\sqrt{2m}} \max_{0=u_0\le u_1\le\cdots\le u_{m-1}\le u_m=1} \min\left(-\frac{1}{m}\sum_{i=1}^{m} B_1^{(i)}(1) + \sum_{i=1}^{m}\left(B_1^{(i)}(u_i) - B_1^{(i)}(u_{i-1})\right), \right.$$
$$\left. -\frac{1}{m}\sum_{i=1}^{m} B_2^{(i)}(1) + \sum_{i=1}^{m}\left(B_2^{(i)}(u_i) - B_2^{(i)}(u_{i-1})\right)\right), \quad (4.75)$$

and the proof of Theorem 1.1 is over.

5 Concluding Remarks

Let us discuss below some potential extensions to Theorem 1.1 and some questions we believe are of interest.

• From the proof presented above, the passage from two to three or more sequences is clear: the minimum over two Brownian functionals becomes a minimum over three or more Brownian functionals, and such a passage applies to the cases touched upon below.

• It is also clear from the proof developed above, that a theorem for two independent sequences of iid (non-uniform) random variables is also valid. Here is what it should look like:

Let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two sequences of iid random variables with values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_m\}$, a totally ordered finite alphabet of cardinality m and with a common law, i.e., $X_1 \stackrel{\mathcal{L}}{=} Y_1$. Let $p_{\max} = \max_{i=1,2,\ldots,m} \mathbb{P}(X_1 = \boldsymbol{\alpha}_i)$ and let k be the multiplicity of $p_{\max}$. Then,

$$\frac{\mathrm{LCI}_n - np_{\max}}{\sqrt{np_{\max}}} \Longrightarrow$$

$$\max_{0=t_0\leq t_1\leq\cdots\leq t_{k-1}\leq t_k=1} \min\left(\frac{\sqrt{1-kp_{\max}}-1}{k}\sum_{i=1}^{k} B_1^{(i)}(1) + \sum_{i=1}^{k}(B_1^{(i)}(t_i) - B_1^{(i)}(t_{i-1})),\right.$$
$$\left.\frac{\sqrt{1-kp_{\max}}-1}{k}\sum_{i=1}^{k} B_2^{(i)}(1) + \sum_{i=1}^{k}(B_2^{(i)}(t_i) - B_2^{(i)}(t_{i-1}))\right), \tag{5.1}$$

where B_1 and B_2 are two k-dimensional standard Brownian motions defined on $[0,1]$.

So, for instance, if $p_{\max}$ is uniquely attained then the limiting law in (5.1) is the minimum of two centered Gaussian random variables.

Using the sandwiching techniques developed in [HL], an infinite countable alphabet result can also be considered with (5.1).

• The loss of independence inside the sequences, and the loss of identical distributions, both within and between the sequences is challenging. Results for these situations will be presented elsewhere. It should also be noted here that the LCIS problem for two or more random permutations has not been studied either and certainly deserves to be studied.

• The length of the longest increasing subsequence is well known to have an interpretation in percolation theory: Indeed, consider the following directed last-passage percolation model in $\mathbb{Z}_+^2$: let $\Pi_2(n,m)$ be the set of directed paths in $\mathbb{Z}_+^2$ from $(0,0)$ to (n,m) with unit steps going either North or East. Given random variables $\omega_{i,j}$, $i \geq 0, j \geq 1$, and interpreting each $\omega_{i,j}$ as the length of time spent by a path at the vertex (i,j), the last-passage time to (n,m) is given by

$$T_2(n,m) = \max_{\pi\in\Pi_2(n,m)}\left(\sum_{(i,j)\in\pi}\omega_{i,j}\right), \tag{5.2}$$

see [BM] for details. In our random word context, when $X = (X_i)_{1\leq i\leq n}$ is a sequence of *iid* random variables taking their values in a totally ordered finite alphabet $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$

of size m, taking $\omega_{i,j} = \mathbf{1}_{\{X_i=\alpha_j\}}$ and $\omega_{0,j} = 0$, $j \geq 1$, which for each i are *dependent* random variables, the length of the longest increasing subsequence of the random word is equal to the last passage-time $T_2(n, m)$, see [BH].

The longest *common* increasing subsequence LCI_n enjoys a similar percolation theory interpretation, but in $\mathbb{Z}_+^3$. Let $\Pi_3(n, n, m)$ be the set of paths in $\mathbb{Z}_+^3$ from $(0, 0, 0)$ to (n, n, m) taking either unit steps towards the top or steps, of any length in the horizontal plane, but neither parallel to the x- nor to the y-axis, i.e.,

$$\begin{aligned}\Pi_3(n,n,m) \;&:=\; \Big\{(u_1,u_2,\dots,u_{n+m}) \in \left(\mathbb{Z}_+^3\right)^{n+m} : u_1=(0,0,1), u_{n+m}=(n,n,m)\\ &\qquad u_{j+1}-u_j \in \big\{(0,0,1),(a,b,0) \text{ with } a,b\in\mathbb{N}\setminus\{0\}\big\}\Big\}.\end{aligned}$$

Given weights $\omega_{i,j,k}$, $i \geq 0, j \geq 0, k \geq 1$, on the lattice, we can consider a quantity analogous to $T_2(n, m)$ in (5.2), namely,

$$T_3^{(c)}(n,n,m) := \max_{\pi\in\Pi_3(n,n,m)} \left(\sum_{(i,j,k)\in\pi} \omega_{i,j,k}\right).$$

In the random word context, taking $\omega_{i,j,k} = \mathbf{1}_{\{X_i=\alpha_k=Y_j\}}$ and $\omega_{0,0,j} = 0$, $j \geq 1$, as weights, gives $\mathrm{LCI}_n = T_3^{(c)}(n, n, m)$.
Note that when $X = Y$, $T_3^{(c)}(n, n, m)$ recovers $T_2(n, m)$ because $T_2(n, m)$ is unchanged if, in (5.2), $\Pi_2(n, m)$ is replaced by

$$\begin{aligned}\widetilde{\Pi}_2(n,m) \;&:=\; \Big\{(u_1,u_2,\dots,u_{n+m}) \in \left(\mathbb{Z}_+^2\right)^{n+m} : u_1=(0,\dots,0,1), u_{n+m}=(n,m)\\ &\qquad u_{j+1}-u_j \in \big\{(0,1),(a,b,0) \text{ with } a,b\in\mathbb{N}\setminus\{0\}\big\}\Big\}.\end{aligned}$$

More generally, for $p \geq 3$ sequences of letters $X^{(j)} = (X_i^{(j)})_{1\leq i\leq n}$, $1 \leq j \leq p$, we can similarly consider

$$\begin{aligned}\Pi_{p+1}(n,\dots,m) := \Big\{&(u_1,u_2,\dots,u_{n+m}) \in \left(\mathbb{Z}_+^{p+1}\right)^{n+m}: u_1=(0,\dots,0,1), u_{n+m}=(n,\dots,n,m)\\ &u_{j+1}-u_j \in \big\{(0,0,1),(a_1,\dots,a_p,0) \text{ with } a_i\in\mathbb{N}\setminus\{0\}\big\}, \forall 1\leq i\leq p\Big\}\end{aligned}$$

and

$$T_p^{(c)}(n,\dots,n,m) := \max_{\pi\in\Pi_{p+1}(n,\dots,n,m)} \left(\sum_{(i_1,\dots,i_p,k)\in\pi} \omega_{i_1,\dots,i_p,k}\right),$$

and observe that the longest common increasing subsequence LCI_n of the p sequences is $\mathrm{LCI}_n = T_p^{(c)}(n, \dots, n, m)$, when $\omega_{i_1,\dots,i_p,k} = \mathbf{1}_{\{X_{i_1}=\dots=X_{i_p}=\alpha_k\}}$ and $\omega_{0,\dots,0,k} = 0$, $j \geq 1$.

• Starting with [Bar] and [GTW] (see, also [BGH], for a further description and up to date references) a strong interaction has been shown to exist between Brownian functionals and maximal eigenvalues of Gaussian random matrices. Likewise, we hypothesize that the max/min functionals obtained here do enjoy a similar strong connection (which might extend to spectra and Young diagrams). Could it be that the right-hand side of (1.1) (with or without the linear terms) has the same law as the maximal eigenvalue of a random matrix model? Even in the binary case, it would be interesting to find the law of the processes $\left(\sqrt{2}\max_{0\le t\le 1}\min(B_1(t)-B_1(1)/2, B_2(t)-B_2(1)/2)\right)_{t\ge 0}$ and $\left(\max_{0\le t\le 1}\min(B_1(t), B_2(t))\right)_{t\ge 0}$ where, say, B_1 and B_2 are two independent standard linear Brownian motions. Very preliminary work on these problems was started with Marc Yor, before his untimely death, and this text is dedicated to his memory.

A Appendix

The purpose of this Appendix is to provide some missing steps in the proof of the main theorem in [HLM] devoted to the binary case as well as to correct the errors present there. The notations and numbering are as in [HLM]. In particular, recall that N_1 (resp. N_2) is the number of zeros in $X_1, \dots, X_n$ (resp. $Y_1, \dots, Y_n$).

Proof of (13). Recall again from [HLM] that

$$\begin{aligned} V_n &= \max_{0\le k\le N_1\wedge N_2}\left(\bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{k}{n}\right)\right)\right) \\ X_n &= \max_{0\le t\le \frac{1}{2}}\left(\bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i(t)\right)\right). \end{aligned}$$

Clearly,

$$X_n \ge \bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{1}{2}\right)\right) = \frac{1}{2}\bigwedge_{i=1,2}\widehat{B}_n^i\left(\frac{1}{2}\right), \tag{A.1}$$

and denote by i_* the index for which the minimum in (A.1) is attained.

Next, if $N_1\wedge N_2 \le n/2$, then $V_n \le X_n$; and similarly if the maximum defining V_n is attained at some $k^* \le n/2$, then $V_n \le X_n$. Otherwise, $N_1\wedge N_2 \ge n/2$ with, moreover, the maximum defining V_n attained at $k^*\in[n/2, N_1\wedge N_2]$ and so:

$$V_n = \bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{k^*}{n}\right)\right).$$

Now, via (A.1),

$$\begin{aligned} V_n - X_n &\le \bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{k^*}{n}\right)\right) - \bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{1}{2}\right)\right) \\ &\le \left(-\frac{1}{2}\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)+\widehat{B}_n^{i_*}\left(\frac{k^*}{n}\right)\right) - \left(-\frac{1}{2}\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)+\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)\right) \\ &= \widehat{B}_n^{i_*}\left(\frac{k^*}{n}\right) - \widehat{B}_n^{i_*}\left(\frac{1}{2}\right) \end{aligned}$$

$$\begin{aligned} &\le \max_{t\in\left[\frac{1}{2},\frac{N_{i_*}}{n}\right]}\left(\widehat{B}_n^{i_*}(t)-\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)\right)\\ &\le \bigvee_{i=1,2}\max_{t\in\left[\frac{1}{2},\frac{N_i}{n}\right]}\left(\widehat{B}_n^{i}(t)-\widehat{B}_n^{i}\left(\frac{1}{2}\right)\right). \end{aligned}$$

Inequality (A.3) replacing (15) of [HLM] and its proof. If $N_1\wedge N_2\ge n/2$, then $X_n\le V_n$ and similarly if the maximum defining X_n is attained for some $t\le (N_1\wedge N_2)/n$, then $X_n=V_n$. Therefore, the remaining case in comparing X_n and V_n consists in $N_1\wedge N_2\le n/2$ and a maximum defining X_n attained at some $t^*\in\left[(N_1\wedge N_2)/n,1/2\right]$. In this case,

$$X_n=\bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i(t^*)\right),$$

and

$$V_n\ge\bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{N_1\wedge N_2}{n}\right)\right). \tag{A.2}$$

Again, denote by i_* the index for which the minimum in (A.2) is attained. Then,

$$\begin{aligned} X_n-V_n &\le \bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i(t^*)\right)-\bigwedge_{i=1,2}\left(-\frac{1}{2}\widehat{B}_n^i\left(\frac{1}{2}\right)+\widehat{B}_n^i\left(\frac{N_1\wedge N_2}{n}\right)\right)\\ &\le \left(-\frac{1}{2}\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)+\widehat{B}_n^{i_*}(t^*)\right)-\left(-\frac{1}{2}\widehat{B}_n^{i_*}\left(\frac{1}{2}\right)+\widehat{B}_n^{i_*}\left(\frac{N_1\wedge N_2}{n}\right)\right)\\ &= \widehat{B}_n^{i_*}(t^*)-\widehat{B}_n^{i_*}\left(\frac{N_1\wedge N_2}{n}\right)\\ &\le \max_{t\in\left[\frac{N_1\wedge N_2}{n},\frac{1}{2}\right]}\left(\widehat{B}_n^{i_*}(t)-\widehat{B}_n^{i_*}\left(\frac{N_1\wedge N_2}{n}\right)\right)\\ &\le \bigvee_{i=1,2}\max_{t\in\left[\frac{N_1\wedge N_2}{n},\frac{1}{2}\right]}\left(\widehat{B}_n^{i}(t)-\widehat{B}_n^{i}\left(\frac{N_1\wedge N_2}{n}\right)\right). \end{aligned} \tag{A.3}$$

Since (15) of [HLM] has to be replaced by (A.3), instead of (16) of [HLM], we now have to prove that for $i=1,2$:

$$\max_{t\in\left[\frac{N_1\wedge N_2}{n},\frac{1}{2}\right]}\left(\widehat{B}_n^{i}(t)-\widehat{B}_n^{i}\left(\frac{N_1\wedge N_2}{n}\right)\right)\xrightarrow{\mathbb{P}}0. \tag{A.4}$$

The difference with (16) of [HLM] is that N, therein, is now replaced by $N_1\wedge N_2$ which is now more complex since one of the two quantities N_1 or N_2 is not independent of $\widehat{B}_n$. To prove (A.4), and so as not to further burden the notation, the superscript i in the Brownian approximation $\widehat{B}_n^i$ is dropped. First, let

$$C_n^1=\left\{\left|N_1-\frac{n}{2}\right|\le\sqrt{n}\ln n\right\},$$

and, in a similar fashion, define C_n^2 by replacing N_1 with N_2. Clearly, $\lim_{n\to+\infty}\mathbb{P}\big((C_n^1)^c\big) = \lim_{n\to+\infty}\mathbb{P}((C_n^2)^c) = 0$. Next, for $\varepsilon > 0$, let

$$A_n = \left\{ \max_{t\in\left[\frac{N_1\wedge N_2}{n},\frac{1}{2}\right]} \left| \widehat{B}_n(t) - \widehat{B}_n\left(\frac{N_1\wedge N_2}{n}\right)\right| \geq \varepsilon \right\}.$$

Then,

$$\mathbb{P}(A_n) \;\leq\; \mathbb{P}\big(A_n\cap C_n^1\cap C_n^2\big) + \mathbb{P}\big((C_n^1)^c\big) + \mathbb{P}\big((C_n^2)^c\big), \tag{A.5}$$

and since on C_n^1 (resp. C_n^2), $N_1 \geq n/2 - \sqrt{n}\ln n$ (resp. $N_2 \geq n/2 - \sqrt{n}\ln n$),

$$\mathbb{P}(A_n\cap C_n^1\cap C_n^2) \;\leq\; \mathbb{P}\left(\left\{\max_{k=\frac{n}{2}-\sqrt{n}\ln n,\ldots,\frac{n}{2}} \left|\sum_{j=N_1\wedge N_2}^{k}\xi_j\right| \geq \varepsilon\sqrt{2n}\right\}\cap C_n^1\cap C_n^2\right), \tag{A.6}$$

where the random variables ξ_j are iid with mean zero and variance one and assuming that $n/2-\sqrt{n}\ln n$ and $n/2$ are integers (if not replace throughout, the first value by its integer part and the second by its integer part plus one). To deal with (A.6), first note that on $C_n^1\cap C_n^2$, $N_1\wedge N_2 \in \left[\frac{n}{2}-\sqrt{n}\ln n, n\right]$, the right-hand side of (A.6) is clearly upper-bounded by

$$\begin{aligned}
&\mathbb{P}\left(\left\{\max_{\frac{n}{2}-\sqrt{n}\ln n\leq \ell\leq k\leq\frac{n}{2}} \left|\sum_{j=\ell}^{k}\xi_j\right| \geq \varepsilon\sqrt{\frac{n}{2}}\right\}\cap C_n^1\cap C_n^2\right)\\
&\qquad\leq \mathbb{P}\left(\left\{\max_{\frac{n}{2}-\sqrt{n}\ln n\leq k\leq\frac{n}{2}} \left|\sum_{j=k}^{n/2}\xi_j\right| \geq \frac{\varepsilon}{2}\sqrt{\frac{n}{2}}\right\}\cap C_n^1\cap C_n^2\right) &\text{(A.7)}\\
&\qquad\leq \frac{2\ln n}{\varepsilon^2\sqrt{n}}, &\text{(A.8)}
\end{aligned}$$

where the inequality in (A.7) follows from the bound

$$\begin{aligned}
\max_{\frac{n}{2}-\sqrt{n}\ln n\leq \ell\leq k\leq\frac{n}{2}} \left|\sum_{j=\ell}^{k}\xi_j\right| \;&\leq\; \max_{\frac{n}{2}-\sqrt{n}\ln n\leq \ell\leq k\leq\frac{n}{2}}\left(\left|\sum_{j=k}^{n/2}\xi_j\right| + \left|\sum_{j=\ell}^{n/2}\xi_j\right|\right)\\
&\leq\; 2\max_{\frac{n}{2}-\sqrt{n}\ln n\leq k\leq\frac{n}{2}}\left|\sum_{j=k}^{n/2}\xi_j\right|,
\end{aligned}$$

while the one in (A.8) is Kolmogorov's maximal inequality. Therefore, the right-hand side of (A.6) converges to zero, finishing, via (A.5), the proof of (A.4). $\square$

References

[Bar] Y. Baryshnikov. *GUEs and queues*, Probab. Theory Relat. Fields vol. 119, pp. 256–274, 2001.

[BGH] F. Benaych-Georges, C. Houdré. *GUE minors, maximal Brownian functionals and longest increasing subsequences in random words*, Markov Processes Relat. Fields, vol. 21, pp. 109-126, 2015.

[BM] T. Bodineau, J. Martin. *A universality property for last-passage percolation paths close to the axis.* Elec. Comm. Prob. vol. 10, pp. 105–112, 2005

[BH] J.-C. Breton, C. Houdré. *Simultaneous asymptotics for the shape of random Young tableaux with growingly reshuffled alphabets.* Bernoulli, vol. 16, no. 2, pp. 471–492, 2010.

[GTW] J. Gravner, C. A. Tracy, H. Widom. *Limit theorems for height fluctuations in a class of discrete space and time growth models*, J. Stat. Phys. vol. 102, pp. 1085–1132, 2001.

[HLM] C. Houdré, J. Lember, H. Maztinger. *On the Longest Common Increasing Binary Subsequence.* C.R. Acad. Sci., Paris Ser. I, vol. 343, pp. 589–594, 2006.

[HL] C. Houdré, T. Litherland. *On the longest increasing subsequence for finite and countable alphabets*, in High Dimensional Probability V: The Luminy Volume (Beachwood, Ohio, USA: Institute of Mathematical Statistics), pp. 185–212, 2009.

[HX] C. Houdré, H. Xu. *On the limiting shape of Young diagrams associated with inhomogeneous random words*, in: High Dimensional Probability VI: The Banff volume Progress in Probability, 66, Birkhauser, pp. 277–302, 2013.

[ITW1] A. Its, C. A. Tracy, H. Widom. *Random words, Toeplitz determinants, and integrable systems. I.* Random matrix models and their applications, pp. 245–258, Math. Sci. Res. Inst. Publ., vol. 40, Cambridge Univ. Press, Cambridge, 2001.

[ITW2] A. Its, C. A. Tracy, H. Widom. *Random words, Toeplitz determinants, and integrable systems. II.* Advances in nonlinear mathematics and science. Phys. D., vol. 152-153, pp. 199–224, 2001.

[Joh] K. Johansson. *Discrete orthogonal polynomial ensembles and the Plancherel measure.* Ann. of Math. (2) 153 (2001), no. 1, 259–296.

[Ker] S. Kerov. *Asymptotic Representation Theory of the Symmetric Group and its Applications in Analysis,* Vol. 219. AMS, Translations of Mathematical Monographs, 2003. (Russian edition: D. Sci thesis, 1994)

[TW] C. A. Tracy, H. Widom. *On the distribution of the lengths of the longest increasing monotone subsequences in random words.* Probab. Theor. Rel. Fields. vol. 119, pp. 350–380, 2001.